Clustering-based Wake Word Detection in Privacy-aware Acoustic Sensor Networks

Timm Koppelmann[1], Luca Becker[1], Alexandru Nelus[1], Rene Glitza[1], Lea Schönherr[2], and Rainer Martin[1]

[1] Institute of Communication Acoustics, Ruhr-Universität Bochum, Bochum, Germany
[2] CISPA Helmholtz Center for Information Security, Saarbrücken, Germany

[1]{firstname.lastname}@rub.de [2]lea.schoenherr@cispa.de

Abstract

This work investigates privacy-aware collaborative wake word detection (WWD) in acoustic sensor networks. To meet state-of-the-art privacy constraints, the proposed WWD scheme is based on privacy-aware unsupervised clustered federated learning that groups microphone nodes w.r.t. active sound sources and on a privacy-preserving high-level feature representation. Using the partition of microphone nodes into clusters, we apply intra- and inter-cluster feature enhancement strategies directly in the privacy-preserving feature domain and thus circumvent the need for communicating privacy-sensitive information between nodes. The approach is demonstrated for an acoustic sensor network deployed in a smart-home environment. We show that the proposed collaborative WWD system clearly outperforms independent decisions of individual microphone nodes.
Index Terms: privacy, wake word detection, clustering, federated learning, unsupervised clustered federated learning

1. Introduction

Legal regulations such as the European Union GDPR [1] define privacy constraints for the processing and storing of personal data, including sensitive speech data that may contain personal or biometric-related information. Privacy concerns are especially valid for smart devices that incorporate microphone sensors and that are interconnected in the form of an acoustic sensor network (ASN) [2]. Their wide-area coverage by means of multiple microphone nodes can help improving many signal processing applications [3] such as source localization [4], event classification [5], and speech enhancement [6] and can be deployed in various environments, some of the most popular being smart-cities [7] and smart-homes [8]. Although ASNs open new perspectives for distributed multi-microphone applications, their interconnection requires privacy-preserving signal sharing and processing approaches in order to comply with legal regulations such as GDPR.

State-of-the-art signal processing in ASNs [3] typically entails that the nodes send raw audio or other information-rich feature-based representations to a centralized party that may also extract additional and potentially private information. Moreover, this data can be subject to interception attacks by eavesdroppers that have infiltrated the network [9]. To reduce the aforementioned privacy risks in the context of WWD, we propose to use a federated learning approach for clustering microphones and to employ a deep neural network (DNN) based privacy-preserving feature representation for WWD. The latter only contains information required by the WWD task while all other information like the content of utterances has been removed. The proposed privacy-preserving WWD scheme is based on our previous work [10], where a privacy-preserving feature representation has been investigated to support the collaboration between a single local node and a more powerful cloud-based server.

Grouping microphone nodes into clusters w.r.t. dominant sound sources may offer significant signal processing advantages [11, 12, 13, 14, 15]. Similarly, WWD performed on the privacy-preserving feature representation may also benefit from feature enhancement techniques. Thus, we focus on the integration of privacy-preserving WWD with a collaborative feature enhancement procedure that makes use of clustered ASN nodes. The employed clustering scheme is based on federated learning (FL) [16] and does not require any raw-signal-based feature representation. Instead it uses DNN weight updates obtained from a light-weight autoencoder deployed at ASN node level as in [14, 15]. The estimation of source-related clusters and WWD in these clusters then allows to confine the operation of any subsequent ASR-based dialogue system [17] to specific spatial locations. We argue that enabling ASR only in clusters that have detected a wake word (in a privacy-aware manner) offers an additional advantage in terms of privacy for the smart-home users [18].

The proposed clustering and collaborative WWD techniques are implemented and validated in a complex smart-home scenario with multiple rooms and open doors that include simultaneously active speech sources and a multitude of ASN nodes. We evaluated the system for very challenging acoustic conditions, where the ASN nodes have to handle low signal-to-noise ratios (SNRs) and several overlapping speakers. We note that the aforementioned WWD approach may be further enhanced by node-based multi-microphone beamforming techniques [19, 20] which are, however, beyond the scope of this paper. Multi-channel signal enhancement approaches using the microphones across multiple nodes require synchronized and possibly non-privacy-preserving signal representations. Therefore, to illustrate the potential of the proposed privacy-aware methods, we assume a single microphone per node only and focus on exploring low-cost collaborative enhancement techniques in the privacy-preserving WWD feature space.

2. UCFL for Clustering ASN Nodes

For clustering ASN nodes, we employ a privacy-aware solution that uses unsupervised clustered federated learning (UCFL) based on Nelus et al. [14, 15]. UCFL is derived from clustered federated learning (CFL) [21, 22, 23] and adapted to the specific case of ASNs. It employs a light-weight autoencoder for each ASN node. Only the autoencoder's bottleneck layer is trained by the ASN nodes. The training is performed for a given number of UCFL communication rounds and minimizes

the mean squared error (MSE) between input and reconstructed log-mel band energy (LMBE) features. The cosine similarity values $A_{i,j}$ of the DNN weight update vectors generated by the training procedure are then used as basis for clustering. The clustering scheme consists of hierarchical bi-partitioning and as such does not require a priori information about the number of sources. UCFL generates a varying number N_C of clusters along with cluster membership values (MVs) μ_i for the nodes n_i of each cluster $c_j \in C$, where C is the set of all clusters. The MVs range from 0 to 1, where $\mu_i = 1$ indicates that node n_i is most representative for the cluster's dominant source. We disregard nodes with low MVs via thresholding, i.e. if $\mu_i \leq v$ then $\mu_i = 0$.

3. Privacy-preserving Wake Word Detection

3.1. WWD model and ASR-based attacker

In the context of this work, we consider multiple nodes n_i that record audio and export a high-level feature representation $\boldsymbol{Z}_i$ which is transmitted to a centralized node for WWD. Similar to our previous work [10], we assume a malicious central processing party or an eavesdropper intercepting those features with the goal of obtaining the transcription related to the recorded audio. In order to suppress the ASR-capability of such an attacker while maintaining a strong WWD performance, we use the privacy-preserving feature extraction implemented in [10] and adapt it to our ASN scenario. We employ a local feature extractor, which is integrated in each node n_i, consisting of a time-delay neural network (TDNN) [24], followed by a bottleneck layer with 16 DNN neurons. The feature extractor transforms mel-frequency cepstral coefficients (MFCCs) into the high-level feature representation $\boldsymbol{Z}_i$. WWD is then performed on a central node using this feature representation, an additional dense layer, and a speech decoder. Furthermore, our WWD system is trained via an adversarial training schedule in order to minimize the attacker's ASR capabilities while maintaining strong WWD performance. Thus, when the attacker performs ASR with intercepted features the word error rate is around 90%.

3.2. Feature fusion

As introduced above, the extracted high-level feature representation $\boldsymbol{Z}_i$ of each node n_i is further processed on a central node in order to perform WWD. Rather than obtaining a WWD decision for each node individually, we implement a high-level *feature fusion* at cluster-level by aggregating over several $\boldsymbol{Z}_i$ within a cluster c_j, where $\mathrm{n}_i \in c_j$:

$$\overline{\boldsymbol{Z}}_j = \frac{1}{\sum\limits_{\mathrm{n}_i \in c_j} \mu_i} \sum_{\mathrm{n}_i \in c_j} \mu_i \boldsymbol{Z}_i. \tag{1}$$

The contribution of each node is controlled via weighting with the related membership values μ_i, optionally constrained by a threshold v as stated in Sect. 2.

3.3. Feature enhancement

Given the envisioned challenging ASN scenario with four simultaneously active sound sources (see Sect. 4.1), we assume that each cluster $c_j \in C$ generated by the UCFL procedure is dominated by a single sound source. As such, the remaining non-dominating sound sources are considered to be interfering (babble) noise for cluster c_j. Considering the privacy constraints introduced in Sect. 1, we further aim to minimize this interference directly in the fused high-level feature space $\overline{\boldsymbol{Z}}_j$. Thus, in addition to feature fusion within one cluster, we propose novel approaches for *high-level feature enhancement* which make use of the auxiliary inter-cluster information derived from UCFL. Even though feature extraction is a nonlinear process in our case, we propose linear methods with the goal of determining the overall feasibility of privacy-preserving feature enhancement. We subsequently present three methods for estimating a noise floor $\hat{\boldsymbol{N}}_j$ at feature level in order to subtract it from the fused features of each target cluster $c_j \in C$:

$$\hat{\boldsymbol{Z}}_j = \alpha \overline{\boldsymbol{Z}}_j - \beta \hat{\boldsymbol{N}}_j, \tag{2}$$

where the impact of the subtraction is controlled via α and β.

3.3.1. Averaging-based estimation

First, we compose the estimation of the noise floor by aggregating high-level features $\overline{\boldsymbol{Z}}_k$ of non-target clusters $c_k \in C \setminus c_j$:

$$\hat{\boldsymbol{N}}_j^{\mathtt{avg}} = \frac{1}{\sum\limits_{c_k \in C \setminus c_j} \psi_{jk}} \sum_{c_k \in C \setminus c_j} \psi_{jk} \overline{\boldsymbol{Z}}_k. \tag{3}$$

Here, the contribution of each single $\overline{\boldsymbol{Z}}_k$ is determined by a specific weighting factor ψ_{jk} which is computed as:

$$\psi_{jk} = \frac{1 - A_{j'k'}}{\sum\limits_{c_l \in C \setminus c_j} 1 - A_{j'l'}}, \tag{4}$$

where $A_{j'k'}$ is the cosine similarity between the weight update vectors of the reference nodes $\mathrm{n}_{j'}$ and $\mathrm{n}_{k'}$ of clusters c_j and c_k generated after the last communication round of UCFL [15]. This is motivated by the fact that a low cosine similarity value indicates that the nodes in the non-target cluster are, most probably, close to an interfering source and thus should have a stronger influence on the noise estimate. Based on preliminary experiments, we set $\alpha = 1.2$ and $\beta = 0.2$ for this approach.

3.3.2. Covariance-based estimation

While $\hat{\boldsymbol{N}}_j^{\mathtt{avg}}$ equally includes all dimensions of feature vectors $\overline{\boldsymbol{Z}}_k$, we now only regard specific information that we assume to be relevant for feature enhancement. Therefore, we consider $\boldsymbol{\xi}_j$ to be the vector of covariance values $\boldsymbol{\xi}_j(d) = \mathtt{cov}\left(\overline{\boldsymbol{Z}}_j(d), \hat{\boldsymbol{N}}_j^{\mathtt{avg}}(d)\right)$ between the d-th element of the target $\overline{\boldsymbol{Z}}_j$ and the previously estimated noise floor vector $\hat{\boldsymbol{N}}_j^{\mathtt{avg}}(d)$. We now define a set D_j^η containing the indices of the η largest elements of $\boldsymbol{\xi}_j$. This is based on the assumption that the feature dimensions corresponding to the largest covariance values are the most representative for the noise-related components in $\hat{\boldsymbol{N}}_j^{\mathtt{avg}}$. Subsequently, we obtain a covariance-based noise floor estimate

$$\hat{\boldsymbol{N}}_j^{\mathtt{cov}}(d) = \begin{cases} \hat{\boldsymbol{N}}_j^{\mathtt{avg}}(d), & d \in D_j^\eta, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Based on initial experiments, we set $\alpha = \beta = 1$ and $\eta = 3$.

3.3.3. MV-based estimation

Finally, we focus on intra-cluster-related information only, assuming that the MV μ_i of each node $\mathrm{n}_i \in c_j$ can be directly related to the amount of interference in the respective audio signal. We now obtain the MV-based estimation

$$\hat{\boldsymbol{N}}_j^{\mathtt{MV}} = \boldsymbol{Z}_{j^*} - \boldsymbol{Z}_{j'}, \tag{6}$$

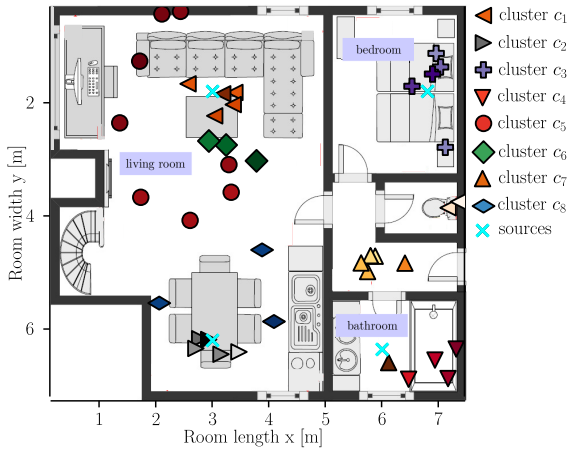

Figure 1: *Floor plan of simulated apartment with cluster estimations for a single scenario. Color intensity is proportional to cluster membership values.*

where $\boldsymbol{Z}_{j*}$ and $\boldsymbol{Z}_{j'}$ indicate the high-level features of the nodes n_{j*} as well as node $\mathrm{n}_{j'}$, with the smallest and largest MV within cluster c_j, respectively. Here we assume that $\boldsymbol{Z}_{j*}$ contains the largest amount of interference related to cluster c_j, while $\boldsymbol{Z}_{j'}$ yields the best representation of the respective dominant sound source for cluster c_j. Again, we set $\alpha = \beta = 1$ here.

4. Experimental Setup

4.1. Scenario description

In order to simulate a challenging ASN scenario we employ the apartment layout from [8] where four simultaneously active sources along with 41 single-microphone ASN nodes are randomly positioned according to the following constraints: two sources in the living room, one in the bedroom and one in the bathroom. At least three nodes are placed within critical distance of each sound source, thus having more direct component energy than reverberation energy. Room impulse responses (RIRs) from all sources to all microphones are then generated using CATT-Acoustic [25] equally to [15]. The room layout is depicted in Fig. 1, including an arbitrarily chosen simulation scenario.

We make use of ten distinct constellations of source-node positions as well as 20 randomly selected gender-balanced speaker groups from our database (see Sect. 4.2), leading to a total of 200 simulation scenarios with a duration of 40 s each. In order to be consistent with [10], we use the wake word "Mister" for all our experiments. During one 40 s simulation scenario, one source is randomly assigned as the wake-word-uttering source, such that the wake word is uttered at least four times. The other three sources are regarded as interference. Given the disproportionately interference, the proposed setup is, thus, very challenging and results in poor SNR conditions (c. f. Sect. 4.3 and 5.1).

4.2. Database description

The dataset used in this work is composed of the *test-clean* subset of the Librispeech corpus [26] and is extended with all utterances from the *train-clean-360* subset that include the wake word "Mister", as described in [10]. In order to match the experimental setup with [15], we remove large parts of silence by applying voice activity detection (VAD) [27] and split every audio file into 10 s segments afterwards. The dataset is further divided into a subset of segments that contain the wake word and a complementary subset of wake-word-free segments. This results in a total of 991 positive examples (occurrences of "Mister") and 77144 negative examples (every other word occurrence) within the scope of the aforementioned 200 simulation scenarios.

4.3. Evaluation measures

4.3.1. UCFL

UCFL is the first processing step in each simulation scenario. We consider fixed source and node positions and do not investigate robustness of the clustering process w.r.t. moving sources and nodes here. UCFL is implemented using the parameters presented in [15] and evaluated using the normalized cluster-to-source distance (CTS) $\tilde{d}_{c_j}^{s_z}$ from source s_z to cluster c_j defined as:

$$\tilde{d}_{c_j}^{s_z} = \frac{\|\rho_{s_z} - \bar{\rho}_{c_j}\|}{\bar{d}_S}, \quad \mathrm{n}_i \in c_j, \tag{7}$$

where ρ_{s_z} is the position of source s_z, $\bar{\rho}_{c_j}$ denotes the geometrical mean of all node positions in the cluster c_j weighted with the respective MVs, and $\bar{d}_S$ is the average of the set of all unique source-pair distances in the simulation [14],[15].

4.3.2. SNR

For the computation of the SNR we consider the signal received from a node's dominant source as the desired signal, while the superposition of the remaining received signals is interpreted as background babble noise. The dominant source of each node is the source towards which the node's cluster exhibits the smallest CTS.

4.3.3. WWD

An extensive description of the proposed WWD system is presented in [10]. The neural network parts are implemented using PyTorch-Kaldi [28] and utilize the Kaldi [29] lattice-decoder. The WWD is deployed in Kaldi by reducing a large-vocabulary language model (LM) so that it contains only information regarding the composition of the wake word "Mister". As evaluation metric, we compute *detection error trade-off* (DET) points between false alarm (FA) and false reject (FR) rates. For a given system, different points are obtained by adjusting the acoustic scaling parameter of the Kaldi decoder. The FA and FR rates are computed for each 10 s segment where a correct wake word detection is assumed only if the number of detected wake words matches the ground truth value in the transcription of the cluster's dominant source. False alarms are only evaluated if the corresponding audio transcription *does not* include a wake word occurrence, while false rejects are evaluated if the transcription *does* include at least one wake word.

5. Results and Discussion

5.1. Clustering performance

UCFL generates a varying number of clusters, with a minimum of four and a maximum of 15 for the considered 200 simulation scenarios. This is dependant on a scenario's reverberation and interference conditions. A typical UCFL result is provided in Fig. 1. Clustering performance is evaluated using the CTS $\tilde{d}_{c_j}^{s_z}$ averaged over 200 simulation scenarios. The results are presented in the left part of Fig. 2 for clusters $c_1 - c_4$ which are the closest clusters corresponding to sources $s_1 - s_4$. These validate the clustering performance since a small CTS value confirms

	c_1	c_2	c_3	c_4
s_1	**0.11**	0.79	0.86	1.0
s_2	0.78	**0.18**	1.2	1.1
s_3	0.8	1.1	**0.17**	1.0
s_4	1.1	1.0	1.0	**0.12**

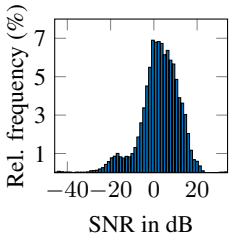

Figure 2: *Normalized cluster-to-source distance (CTS) between clusters $c_1 - c_4$ and sources $s_1 - s_4$, averaged over 200 simulation scenarios (left). Histogram of SNR-values for each ASN node related to the presented clusters $c_1 - c_4$ measured over 200 simulation scenarios (right).*

that a cluster c_j is close to a source s_z, while $\tilde{d}_{c_j}^{s_z} > 1$ implies that the CTS is larger than the average distance between sources [14, 15]. Based on the clustering results we also generate a histogram of the SNR measurements introduced in Sect. 4.3.2. This is displayed in the right part of Fig. 2 and further underlines the challenging acoustic conditions included in the proposed scenario.

5.2. Reference WWD performance

We evaluate different privacy-preserving WWD approaches and present them in Fig. 3. Note that in all cases, in order to distinctly observe the effects of UCFL, the WWD system has only been trained with clean data. We first discuss WWD performance without additional high-level feature enhancement (see Sect. 3.3) starting with the baseline system (blue markers). Here, we do not employ UCFL information and, thus, are not able to apply high-level feature fusion. This leads to poor WWD performance with FR rates higher than 0.51 due to the challenging SNR conditions presented earlier.

Based on preliminary studies, high-level feature fusion is applied for all evaluations using an MV threshold $v = 0.5$ (see Sect. 3.2). Only clusters $c_1 - c_4$ are taken into account as they are the most representative for each source. The effects of high-level feature fusion, without yet applying any feature enhancement, are indicated in Fig. 3 by yellow markers. We observe a significant increase in WWD performance thanks to the incorporation of UCFL information, but, nevertheless, FR rates do not drop below 0.42 thus motivating the need for additional enhancement.

Before further analyzing the effects of feature enhancement we first look at the best achievable WWD performance (green markers) by assuming an ideal case without interference. As such, an upper bound on WWD performance is obtained with FA rates of 0.01 when considering FR rates below 0.12.

5.3. WWD performance with feature enhancement

As presented in Sect. 3.3, we now proceed with three different noise floor estimation methods that allow the subtraction of the latter from the target feature representation. It can be observed in Fig. 3 that all approaches lead to the desired improvement in FR rate, nevertheless, at the expense of increased FA rates.

In general, the results for the three enhancement methods provide very similar performance, reaching FR rates of 0.12 (which is close to the aforementioned upper bound) at FA rates of around 0.125. Compared to other published works (e.g., [30] with FA rates below 0.03 for FR rates around 0.13 in noisy environments) our FA rates appear to be rather large. However,

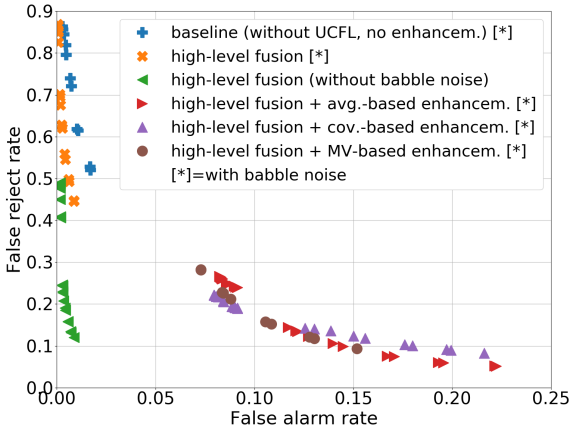

Figure 3: *WWD detection error trade-off for the baseline system without UCFL, approaches applying high-level feature fusion, and additional feature enhancement methods. Green markers serve as an upper bound for WWD performance, excluding any interference.*

given the chosen complex apartment model along with four simultaneously active speakers it becomes obvious that WWD is much more challenging and a reasonable trade-off between FA and FR rates is necessary. To enhance the performance in these difficult acoustic conditions, further improvements of the proposed feature enhancement method as well as node-based adaptive beamformers are desirable.

While solely applying high-level feature fusion (orange markers in Fig. 3) could not prevent the WWD system from rejecting most of the wake word occurrences, it is obvious that further enhancing the high-level feature space using UCFL information can lead to significant improvements. It has also been shown that using inter-node covariance information helps to extract those dimensions of the feature vector that are useful for the estimation of babble-noise related interference. These findings motivate further investigations in new, privacy-preserving speech enhancement methods (including deep learning-based approaches) which may then directly be applied in the high-level feature space.

6. Conclusions

In this work we have shown that a collaborative WWD using UCFL information outperforms a WWD decision solely based individual nodes in a challenging privacy-aware ASN scenario. Nevertheless, a large amount of wake words is still rejected under the bad SNR conditions due to continuous interference from babble noise. Therefore, novel privacy-preserving enhancement methods applied directly in the high-level feature space have shown strong potential for further improving the WWD performance. Given the aforementioned low-SNR conditions in combination with our strict privacy constraints, the presented enhancement approaches were not yet capable of providing a detection error trade-off close to the interference-free upper bound WWD performance. Thus, in future works we will study the presented idea of feature-level interference compensation in conjunction with nonlinear, deep learning based techniques.

7. Acknowledgements

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

8. References

[1] P. Voigt and A. v. d. Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st ed. Springer Publishing Company, Incorporated, 2017.

[2] S. Pasha, C. Ritz, and J. Lundgren, "A survey on ad hoc signal processing: Applications, challenges and state-of-the-art techniques," in *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2019, pp. 1–6.

[3] 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)*. IEEE, 2011, pp. 1–6.

[4] G. F. Miller, A. Brendel, W. Kellermann, and S. Gannot, "Misalignment recognition in acoustic sensor networks using a semi-supervised source estimation method and markov random fields," *arXiv:2011.03432*, 2020.

[5] J. Ebbers, L. Drude, R. Haeb-Umbach, A. Brendel, and W. Kellermann, "Weakly supervised sound activity detection and event classification in acoustic sensor networks," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2019, pp. 301–305.

[6] S. Markovich-Golan, A. Bertrand, M. Moonen, and S. Gannot, "Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks," *Signal Processing*, vol. 107, pp. 4–20, 2015.

[7] M. Cartwright, A. E. M. Mendez, J. Cramer, V. Lostanlen, H.-H. Wu, J. Salamon, and J. Bello. (2019) DCASE 2019: Urban Sound Tagging. Accessed February, 2020. [Online]. Available: http://dcase.community/challenge2019/task-urban-sound-tagging.

[8] G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, and P. Karsmakers, "DCASE 2018 Challenge - Task 5: Monitoring of domestic activities based on multi-channel acoustics," KU Leuven, Tech. Rep., 2018.

[9] G. S. Gaba, G. Kumar, H. Monga, T. Kim, and P. Kumar, "Robust and lightweight mutual authentication scheme in distributed smart environments," *IEEE Access*, vol. 8, pp. 69 722–69 733, 2020.

[10] T. Koppelmann, A. Nelus, L. Schönherr, D. Kolossa, and R. Martin, "Privacy-Preserving Feature Extraction for Cloud-Based Wake Word Verification," in *Proc. Interspeech 2021*, 2021, pp. 876–880.

[11] S. Gergen, A. Nagathil, and R. Martin, "Classification of reverberant audio signals using clustered ad hoc distributed microphones," *Signal Process.*, vol. 107, no. C, p. 21–32, Feb. 2015.

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

[13] 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.

[14] A. Nelus, R. Glitza, and R. Martin, "Estimation of microphone clusters in acoustic sensor networks using unsupervised federated learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 761–765.

[15] ——, "Unsupervised clustered federated learning in complex multi-source acoustic environments," in *29th European Signal Processing Conference, EUSIPCO 2021*. IEEE, 2021.

[16] B. McMahan, E. Moore, D. Ramage, and B. Arcas, "Federated learning of deep networks using model averaging," *arXiv:1602.05629*, 2016.

[17] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistance: Combining signal processing with deep-learning techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, p. 111–124, 2019.

[18] P. P. Zarazaga, S. Das, T. Bäckström, V. V. R. Vegesna, and A. K. Vuppala, "Sound privacy: A conversational speech corpus for quantifying the experience of privacy." in *Interspeech*, 2019, pp. 3720–3724.

[19] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *ICASSP 2018, Calgary, Canada*, 2018.

[20] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel asr system," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[21] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2020.

[22] 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.

[23] 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*. IEEE, 2020, pp. 8861–8865.

[24] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech 2015*. ISCA, 2015, pp. 3214–3218.

[25] B.-I. Dalenbäck, *TUCT v2.0e:1*, CATT, Mariagatan 16A, SE-41471 Gothenburg, Sweden, 2019. [Online]. Available: http://www.catt.se

[26] D. Povey. LibriSpeech ASR corpus. Accessed February, 2020. [Online]. Available: http://www.openslr.org/12.

[27] Google WebRTC. Accessed February, 2021. [Online]. Available: https://webrtc.org.

[28] M. Ravanelli, T. Parcollet, and Y. Bengio, "The pytorch-kaldi speech recognition toolkit," in *Proc. of ICASSP*, 2019.

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

[30] G. Chen, C. Parada, and T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*. IEEE, 2015, pp. 5236–5240.