



Online Learning of Open-set Speaker Identification by Active User-registration

Eunkyung Yoo, Hyeonseop Song, Teahyeong Kim, Chul Lee

AI Lab, CTO Division, LG Electronics, Seoul, Republic of Korea

{eunkyung.ryu, hyeonseop.song, taehyeong.kim, clee.lee}@lge.com

Abstract

Registering each user’s identity for voice assistants is burdensome and complex for multi-user environments like a household scenario. This is particularly true when the registration needs to happen on-the-fly with a relatively minimum effort. Most of the prior works for speaker identification (SID) do not seamlessly allow the addition of new speakers as these do not support *online* updates. To deal with such limitation, we introduce a novel online learning approach to open-set SID that can actively register unknown users in the household setting. Based on MPART (Message Passing Adaptive Resonance Theory), our method performs online active semi-supervised learning for open-set SID by using speaking embedding vectors to infer new speakers and request user’s identity. Our method progressively improves the overall SID performance without forgetting, making it attractive for many interactive real-world applications. We evaluate our model for the online learning setting of an open-set SID task where new speakers are added on-the-fly, demonstrating its superior performance.

Index Terms: open-set speaker identification, online active learning, message passing adaptive resonance theory

1. Introduction

Speaker identification (SID) aims to detect the speaker identity of a given utterance based on user’s unique vocal characteristics. SID is an important feature for personal voice assistant (PVA) services, especially for their household and mobile applications. In most cases, users manually register their voices in advance for these applications. While voice assistants aim to offer the most seamless registration experience, it is often a huddle for new users to register, especially when these are not that familiar with PVAs.

To deal with such limitation, open-set SID approaches [1, 2, 3] have been considered as an alternative to register new users. That is, open-set SID takes into account the possibility that a given utterance does not belong to any of the enrolled speakers. Most of the prior works on open-set SID do not properly address the cases where unknown users can be registered as additional users on-the-fly without involving any offline computation. In this work, we study how to actively register and identify speakers for unspecified users via online learning.

Note that active learning is a relatively inexpensive yet easy way of enabling speaker identification of new users. That is, we could simply register new speakers by actively querying users for utterances of unknown identity. There are still some limitations with utilizing active learning methods that sample from a pool of stored data. First, it is not that easy to identify the speaker’s identity even for humans [4]. Also, there are potential privacy concerns as we need to directly query the speaker’s utterance, which might contain personal information, rather than the speaker embedding vectors [5, 6].

A more attractive approach would be directly requesting the

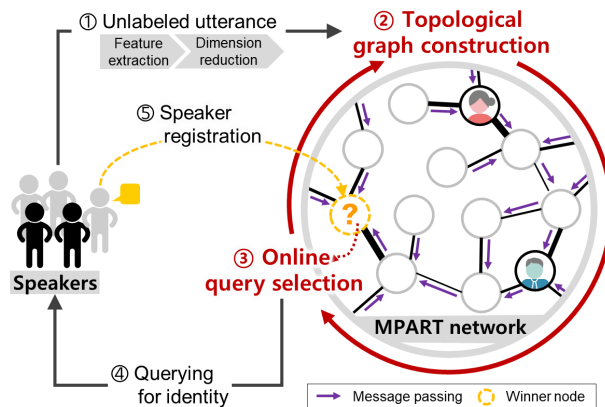


Figure 1: Overview of the proposed method: (1) Extract the speaker embedding vector from the input utterance and reduce its dimensions. (2) Build the topological graph of input samples online. (3) Select a representative and uncertain sample for unknown speaker detection. (4) Query the speaker’s identity. (5) Update the model to reflect the speaker’s label.

user’s identity information without storing any utterance data. When the voice assistant receives the utterance from an unknown speaker, it can immediately ask who is speaking to. Online active learning (OAL) [7, 8] is particularly suitable for this SID scenario in which data labeling cost is prohibitive and user data cannot be stored due to privacy concerns. OAL performs the selection and querying of samples in addition to model updates on-the-fly. To the best of our knowledge, OAL has never been used for open-set SID.

Finally, we can take the full advantage of semi-supervised learning (SSL) to improve the overall performance of our model by additionally utilizing unlabeled data along with a small number of labeled data. In the past, several SSL approaches such as label propagation and graph-based methods have been proposed for SID [9, 10, 11, 12]. Most of them, though, require iterative training in a batch mode, making them difficult to use for the online learning of open-set SID tasks, where the learning targets continuously evolve [13].

Considering all these aspects, we propose a novel online active semi-supervised learning method for open-set SID based on *Message Passing Adaptive Resonance Theory* (MPART) [14]. MPART is an ART-based model [15] that can effectively learn input data distribution (i.e., topological graph) on-the-fly without forgetting. It can also efficiently estimate class labels and uncertainties associated input samples using a small amount of labeled data in a class-incremental learning manner. We take advantage of MPART to query the identities of unknown speakers and register them as additional users for open-set SID. To experimentally validate our approach, we design an online learning task for open-set SID where the storage of utterances is

not allowed and a set of new speakers can be added at any time. In our experiments that only allow the access to a small amount of identity labels, the results show that our proposed method significantly outperforms other baselines. We illustrate an overview of our proposed method in Figure 1.

2. Related Works

2.1. Open-set speaker identification

Studies on open-set SID include implementing an adaptive-Gaussian mixture model (GMM) with a fast scoring technique [16] and showing the effectiveness of the i-vector and GMM-UBM methods [17]. Recently, a challenge on discriminating blacklist [18] has been presented, and several methods have been proposed [3, 19] to solve this task. Most of the open-set SID studies do not address online learning of additional speakers' identities. As an online learning method, a framework was introduced to train SID model on the fly without prior registration [20]. However, this method is based on the iterative feedback from users which does not utilize active user-registration.

2.2. Active learning on speaker recognition

Some studies are introduced neural network based active learning methods to compare performance with conventional approaches (e.g., GMM) for open-set SID [21, 22]. In the field of speaker diarization [23], there are studies to reduce the cost and lengthy process of manual annotation [24] and improve performance through the active query [25]. There are researches on speaker recognition that compare performance using an uncertain sample query strategy with various audiovisual datasets [26] and develop the technique to maximize cluster size and manifold coverage while minimizing the number of queries [27]. Although these methods contributed to reducing the need for labeled data, they require iterative data access for query selection, making them unsuitable for online learning.

3. Methods

3.1. Problem settings and overview

The proposed method aims to classify speakers from the input data stream of utterances without prior knowledge such as pre-registered speakers or total number of users. Formally, for input utterance stream $X = \{x_1, x_2, \dots, x_t\}$, we train a model online and predict the speaker's identity label $Y = \{y_1, y_2, \dots, y_t\}$. All utterances x_t are given without labels, and the model can infer unknown speakers and query the identity y_t of the user to improve SID performance.

To this end, we first extract speaker embedding vector (i.e., x-vector [28]) h_t from the utterance x_t , and the reduced n_r -dimensional feature vector $r_t \in [0, 1]^{n_r}$ is obtained from h_t using the pre-trained Parametric-UMAP [29]. Then, as proposed in MPART [14], we gradually build a topological graph by learning the distribution and correlations of the input samples. In the topological graph, nodes represent categories of input samples, and edges indicate the similarity between the nodes. We can label each node and perform classification using the label density of the node to which the input sample corresponds (i.e., the winner node). To compensate for the lack of labeled nodes, we use the message passing method on the graph to classify the speakers and query unknown users for their identity. All processes are performed online, and the overall process is described in Algorithm 1. The following sections briefly describe

MPART's methods for topological graph construction, speaker identification, and active querying.

3.2. Topological graph construction

To form the nodes of the topological graph, complemented code $I_t = [r_t, \bar{1} - r_t]$ is used in input layer, as in Fuzzy ART [30]. All nodes j are connected to the input layer with adaptive weights w_j , and are created or activated based on matching function M_j and choice function T_j defined in Equation 1.

$$M_j(I_t) = \frac{\|I_t \wedge w_j\|_1}{\|I_t\|_1}, \quad T_j(I_t) = \frac{\|I_t \wedge w_j\|_1}{\alpha + \|w_j\|_1} \quad (1)$$

where \wedge denotes element-wise minimum operation, $\|\cdot\|_1$ is L1 normalization and $\alpha > 0$ is a hyperparameter for the choice function. Input I_t is compared with all nodes to get $M_j(I_t)$. To become candidates of winner, $M_j(I_t)$ is greater than or equal to a vigilance parameter $\rho \in [0, 1]$. The winner node J_t is selected by the largest value $T_j(I_t)$ among candidates, and the remaining nodes are referred to co-activated nodes.

The winner node is updated with a learning rate $\beta \in [0, 1]$ and raise the winning count d_{J_t} by Equation 2. If there is no winner, a new node J_t is created and initialized to $w_{J_t} = I_t$ and $d_{J_t} = 1$.

$$w_{J_t}^{new} = \beta(I_t \wedge w_{J_t}^{old}) + (1 - \beta)w_{J_t}^{old} \\ d_{J_t}^{new} = d_{J_t}^{old} + 1 \quad (2)$$

With the formation of nodes, if multiple nodes are activated together, the co-activated counts $c_{J_t v}$ between winner node J_t and co-activated nodes $v \neq J_t$ are increased by 1. The edge weight e_{ij} of the topological graph is defined as Equation 3.

$$e_{ij} = c_{ij} / (d_i + d_j) \quad (3)$$

where c_{ij} is the co-activated count of nodes i and j . The edge weight e_{ij} is always in between 0 and 1, so it can be used for message passing on the graph without normalization.

3.3. Speaker identification and active querying

3.3.1. Message passing

The message passing method for MPART is defined as Equation 4 for node classification and active querying.

$$X_i^{(l)} = X_i^{(l-1)} + \delta \sum_{j \in \mathcal{N}_i} e_{ij} X_j^{(l-1)}, \quad \forall i \in \mathcal{N}_{J_t}^{(0:L-l)} \quad (4)$$

where $\delta \in [0, 1]$ is a hyperparameter for propagation rate. X_i and X_j are information vectors such as label density and winning count, and \mathcal{N}_i is the set of all neighbors of node i . This method is used repeatedly on multiple layers to aggregate a broader range of information. Finally, we can use the node information $X_i^{(L)}$ of the final layer L to perform the task we want.

3.3.2. Speaker Identification

The speaker identification of the input sample x_t is done by estimating the class label of the winner node J_t . When a label y_t is received to winner node, the label density $q_{J_t}(y_t)$ is increased by 1. The class of a node can be evaluated not only by the labels of the node which the sample belongs to, but also the labels of the surrounding nodes even in the case where a rare label is given. The class probability distribution $p_t(y)$ and the

estimated speaker \hat{y} of input x_t is obtained using the aggregated label density $q_{J_t}^{(L)}$ as shown in equation 5.

$$p_t(y) = \frac{q_{J_t}^{(L)}(y)}{\sum_{y' \in C} q_{J_t}^{(L)}(y')}, \quad \hat{y} = \arg \max_{y' \in C} p_t(y') \quad (5)$$

where C is a set of known speaker's labels.

3.3.3. Active querying

We first use the aggregated winning count $d_{J_t}^{(L)}$ of the winner node J_t to select representative samples for SID. The aggregated winning count $d_{J_t}^{(L)}$ increases as the number of input samples that activate the winner node J_t and its surrounding nodes increases. Therefore, it tends to have large values at the center of the distribution of feature vectors for a given speaker. We define the *density score* s_t of the input sample x_t using $d_{J_t}^{(L)}$ as in Equation 6, where $k_d > 0$ is a constant for sensitivity. By selecting samples where s_t is greater than the *density threshold* $\theta_d \in [0, 1)$, we can query representative samples.

$$s_t = \tanh(k_d \cdot d_{J_t}^{(L)}) \quad (6)$$

We also utilize uncertainty u_t , which can be seen as an epistemic uncertainty [31], using quantitative information of label density $q_{J_t}^{(L)}$ of winner node J_t as shown in Equation 7.

$$u_t = 1 - \tanh\left(k_u \sum_{y \in C} q_{J_t}^{(L)}(y)\right) \quad (7)$$

where $k_u > 0$ is the sensitivity constant for u_t . Uncertainty u_t has a high value in the label-poor regions of the input data distribution. We can query informative samples by choosing samples with u_t greater than the *uncertainty threshold* θ_u . Unlike the originally proposed method [14], which used *density-weighted query selection score*, we utilize density score s_t and uncertainty u_t respectively for representative and informative query selection. Concretely, the model query samples that satisfy both $s_t > \theta_d$ and $u_t > \theta_u$ to acquire labels and incrementally improve SID performance.

4. Experiments

4.1. Implementation details

We used the VoxCeleb1 [32] and VoxCeleb2 [33] datasets containing a large scale of human speech for the pre-training of feature extractor and SID tasks. VoxCeleb1 and VoxCeleb2 consists of 1,251 and 6,112 speakers with 153,516 and 1,128,246 utterances, respectively. We first trained a speaker embedding extractor [6, 28, 34] using VoxCeleb2 dataset. In this process, the MUSAN [35] and RIR [36] datasets were used together for data augmentation. Then, we trained Parametric-UMAP for dimensionality reduction using the extracted embedding vectors (i.e., 512-dim x-vector) from 30% of the VoxCeleb2 dataset. For Parametric-UMAP training, cosine similarity was used as a distance function, and finally, the utterance samples were embedded as the 5-dimensional feature vectors.

The online learning tasks of open-set SID was performed using the VoxCeleb1 dataset. We sifted 540 speakers with more than 100 utterances on VoxCeleb1. For each speaker, 10 utterances were randomly selected for evaluation, and the rest were used for training. We constructed a validation dataset of 10 households consisting of random speakers and set the hyper-parameters $\rho=0.96$, $\beta=0.5$, $\delta=0.7$, and $L=4$ empirically.

Algorithm 1: Open-set SID using MPART.

```

1  $V \leftarrow \{\}, C \leftarrow \{\}$ 
2 for  $x_t$  in input data stream do
3    $h_t \leftarrow \text{FeatureExtraction}(x_t)$ 
4    $r_t \leftarrow \text{DimensionReduction}(h_t)$ 
5    $I_t \leftarrow [r_t, \vec{1} - r_t]$ 
6    $A \leftarrow \{\}$ 
7   for  $j$  in  $1, \dots, |V|$  do
8      $M_j \leftarrow \|I_t \wedge w_j\|_1 / \|I_t\|_1$ 
9      $T_j \leftarrow \|I_t \wedge w_j\|_1 / (\alpha + \|w_j\|_1)$ 
10    if  $M_j \geq \rho$  then
11       $A \leftarrow A \cup \{j\}$ 
12  if  $A$  is empty then
13     $J_t \leftarrow |V| + 1, V \leftarrow V \cup \{J_t\}$ 
14     $c_{J_t v} \leftarrow 0, c_{v J_t} \leftarrow 0 \quad \forall v \in V - \{J_t\}$ 
15     $q_{J_t}(y) \leftarrow 0 \quad \forall y \in C$ 
16     $w_{J_t} \leftarrow I_t, d_J \leftarrow 1$ 
17  else
18     $J_t \leftarrow \arg \max_{j \in A} (T_j)$ 
19     $c_{J_t v} \leftarrow c_{J_t v} + 1 \quad \forall v \in A - \{J_t\}$ 
20     $c_{v J_t} \leftarrow c_{v J_t} + 1 \quad \forall v \in A - \{J_t\}$ 
21     $w_{J_t} \leftarrow \beta(I_t \wedge w_{J_t}) + (1 - \beta)w_{J_t}$ 
22     $d_{J_t} \leftarrow d_{J_t} + 1$ 
23   $q_{J_t}^{(L)}, d_{J_t}^{(L)} \leftarrow \text{MessagePassing}(J_t, c, d, q)$ 
24   $p_t, \hat{y} \leftarrow \text{SpeakerIdentification}(q_{J_t}^{(L)})$ 
25   $s_t \leftarrow \text{DensityScoreEstimation}(d_{J_t}^{(L)})$ 
26   $u_t \leftarrow \text{UncertaintyEstimation}(q_{J_t}^{(L)})$ 
27  if  $s_t > \theta_d$  and  $u_t > \theta_u$  then
28     $y_t \leftarrow \text{QueryLabel}(x_t)$ 
29     $q_{J_t}(y_t) \leftarrow q_{J_t}(y_t) + 1$ 
30     $C \leftarrow C \cup \{y_t\}$ 

```

4.2. Experimental setting

To investigate our approach, we designed two tasks in which data is given as a stream and not stored. The *Task1* aims to evaluate the performance of the online active learning method for open-set SID. We also experimented with *Task2* to verify whether performance degrades when the number of speakers increases. In each experiment, we set up a household with S speakers randomly selected from a pool of 540 speakers. SID accuracy was averaged over 1,000 households to see statistically significant results.

Task1. All utterances are given in random order in a household consisting of various speaker numbers $S = \{4, 6, 8\}$. We compared our method to SSL-baselines of 'Person' and 'Random', trained with N pre-labeled samples. Additionally, the effect of the number of labels on SID performance is reported.

- **Person:** This method simulates the case in which the speakers individually register their utterances. We provided a total of N labeled utterances with equal numbers for each speaker. Since the labeled samples of all speakers are acquired evenly, it is advantageous for SID.
- **Random:** N labeled samples are randomly given for the utterances of a household. In this method, the model randomly queries the user's identity without considering the distribution of the data and unknown speakers. Hence, it

is a baseline for whether our method detects unknown speakers well and queries beneficial samples.

- **Ours:** Although no pre-labeled data is provided, the model can query the user’s identity using the proposed method. For the comparison to SSL-baselines, we set the parameters for Q as Ours-1 ($\theta_d = 0.96, \theta_u = 0.96$) and Ours-2 ($\theta_d = 0.92, \theta_u = 0.80$).

Task2. In this task, we evaluated the performance of class incremental learning (CIL) in SID. CIL suffers from the *catastrophic forgetting* which loses the previous learned information when new data is learned. We verified whether this phenomenon occurs by evaluating the SID performance when the utterances of new speakers are additionally learned in the proposed model. To this end, we divided the total number of speakers per household $S = \{6, 8\}$ into two groups (i.e., ‘Group-A’ and ‘Group-B’). Then, the SID performance of the model, which learned the utterances of ‘Group-A’ first and additionally learned the utterances of ‘Group-B’, was evaluated for all speakers in the household. Finally, we compared the results of *Task2* with *Task1*.

Table 1: Comparison of open-set SID accuracy (mean \pm std) between our model and SSL-baseline. ‘# Speaker’ refer to the number of speakers in household. ‘N/S’ or ‘Q/S’ denotes the number of labeled data or queries per person.

# Speaker	Method	N/S (or Q/S)	Accuracy (%)
4	Person	2	90.0 \pm 10.4
		3	91.1 \pm 9.6
	Random	2	77.5 \pm 14.9
		3	85.4 \pm 13.0
	Ours-1	1.93	92.2 \pm 8.7
	Ours-2	3.05	92.8 \pm 8.1
6	Person	2	85.2 \pm 9.6
		3	87.4 \pm 8.7
	Random	2	72.8 \pm 12.5
		3	80.1 \pm 11.3
	Ours-1	1.87	87.4 \pm 9.5
	Ours-2	3.04	88.6 \pm 8.0
8	Person	2	80.9 \pm 9.3
		3	83.7 \pm 8.4
	Random	2	69.0 \pm 11.5
		3	76.3 \pm 10.5
	Ours-1	1.86	83.8 \pm 8.7
	Ours-2	3.00	85.2 \pm 8.1

Table 2: Open-set SID accuracy (mean \pm std) on the CIL environments. ‘S’ + ‘S’ indicates the number of speakers in the first added group and the later joined group. ‘Q/S’ denotes the number of queries per person.

# Speaker	Method	Q/S	Acc (%)
3 + 3	Ours-1	1.92	89.67 \pm 8.0
	Ours-2	3.05	90.83 \pm 7.2
4 + 4	Ours-1	1.89	86.23 \pm 7.8
	Ours-2	3.05	87.43 \pm 7.4

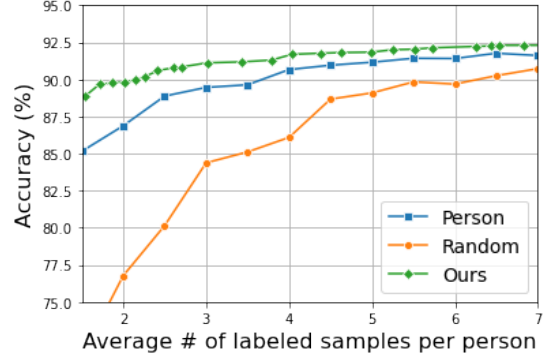


Figure 2: Comparison of open-set SID accuracy with the labeled sample increases. The household consisted of 4 speakers.

4.3. Experimental results

Table 1 summarizes the performance of the proposed method and SSL-baselines. Our method outperforms both SSL-baseline with ‘Person’ and ‘Random’ when Q is almost equal to N . The proposed method achieves 1.95% and 11.48% improvement against the SSL-baselines with ‘Person’ and ‘Random’, respectively. It is clear from these results that the proposed method on active query could select more important samples because the uncertainty prediction and sample selection have reflected the distribution of continuous input sequences. We observed the proposed method also has better stability through small standard deviation accuracy.

Figure 2 depicts the identification accuracy as the number of queries increases to demonstrate the role of active querying. The performance of the proposed method (‘Ours’) is better than the baselines, and the active query is more effective when there are fewer labels.

The results of *Task2* where new speaker groups are added separately are shown in Table 2. Most parametric models trained in batch mode suffer from catastrophic forgetting in incremental learning environments such as *Task2*. On the other hand, the proposed method showed slightly higher performance in *Task2* than in *Task1* without forgetting. This result is because there is an advantage in selecting representative samples for each speaker when the utterances are input in speaker order rather than random.

5. Conclusions

In this work, we introduce an online learning method that can actively register user in open-set SID tasks. The proposed method is based on MPART for online active semi-supervised learning, which can effectively detect unregistered users using limited labeled data and query the user’s identity. We evaluated our method on tasks where the number of users is unknown in advance and data is given as a stream, showing that our approach has significant advantages in open-set SID applications. Over the past decades, AI techniques have been remarkable strides, however, some people are alienated from its benefits. We believe that our research will enhance the equal opportunity of speech technology for our society by providing a natural interaction with machines.

6. References

- [1] R. Karadaghi, “Open-set speaker identification,” 2018.
- [2] S. Imoscopi, V. Grancharov, S. Sverrisson, E. Karlsson, and H. Pobloth, “Experiments on open-set speaker identification with discriminatively trained neural networks,” *arXiv preprint arXiv:1904.01269*, 2019.
- [3] K. Wilkinghoff, “On open-set speaker identification with ivectors,” in *The Speaker and Language Recognition Workshop (Odyssey)*. ISCA, 2020, pp. 408–414.
- [4] N. K. Sharma, S. Ganesh, S. Ganapathy, and L. L. Holt, “Talker change detection: A comparison of human and machine performance,” *The Journal of the Acoustical Society of America*, vol. 145, no. 1, pp. 131–142, 2019.
- [5] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.
- [6] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.
- [7] E. Lughofer, “On-line active learning: A new paradigm to improve practical useability of data stream modeling methods,” *Information Sciences*, vol. 415, pp. 356–376, 2017.
- [8] D. Liu, P. Zhang, and Q. Zheng, “An efficient online active learning algorithm for binary classification,” *Pattern Recognition Letters*, vol. 68, pp. 22–26, 2015.
- [9] N. Fazakis, S. Karlos, S. Kotsiantis, and K. Sgarbas, “Speaker identification using semi-supervised learning,” in *International Conference on Speech and Computer*. Springer, 2015, pp. 389–396.
- [10] L. Chen, V. Ravichandran, and A. Stolcke, “Graph-Based Label Propagation for Semi-Supervised Speaker Identification,” in *Proc. Interspeech 2021*, 2021, pp. 4588–4592.
- [11] J. Wang, X. Xiao, J. Wu, R. Ramamurthy, F. Rudzicz, and M. Brudno, “Speaker attribution with voice profiles by graph-based semi-supervised learning,” *arXiv preprint arXiv:2102.03634*, 2021.
- [12] M. Yamada, M. Sugiyama, and T. Matsui, “Semi-supervised speaker identification under covariate shift,” *Signal Processing*, vol. 90, no. 8, pp. 2353–2361, 2010.
- [13] T. Kim, I. Hwang, G.-C. Kang, W.-S. Choi, H. Kim, and B.-T. Zhang, “Label propagation adaptive resonance theory for semi-supervised continuous learning,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4012–4016.
- [14] T. Kim, I. Hwang, H. Lee, H. Kim, W.-S. Choi, J. J. Lim, and B.-T. Zhang, “Message passing adaptive resonance theory for online active semi-supervised learning,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 5519–5529.
- [15] S. Grossberg, “Competitive learning: From interactive activation to adaptive resonance,” *Cognitive science*, vol. 11, no. 1, pp. 23–63, 1987.
- [16] J. Fortuna, P. Sivakumaran, A. Ariyaeeinia, and A. Malegaonkar, “Open-set speaker identification using adapted gaussian mixture models,” in *Ninth European conference on speech communication and technology*, 2005.
- [17] R. Karadaghi, H. Hertlein, and A. Ariyaeeinia, “Effectiveness in open-set speaker identification,” in *2014 International Carnahan Conference on Security Technology (ICCST)*. IEEE, 2014, pp. 1–6.
- [18] S. Shon, N. Dehak, D. Reynolds, and J. Glass, “Mce 2018: The 1st multi-target speaker detection and identification challenge evaluation,” *arXiv preprint arXiv:1904.04240*, 2019.
- [19] E. Khoury, K. Lakhndhar, A. Vaughan, G. Sivaraman, and P. Nagarsheth, “Pindrop labs’ submission to the first multi-target speaker detection and identification challenge,” in *INTERSPEECH*, 2019, pp. 1502–1505.
- [20] B. Lin and X. Zhang, “Voiceid on the fly: A speaker recognition system that learns from scratch,” in *INTERSPEECH*, 2020.
- [21] L. Wang, K. Chen, and H. S. Chi, “Towards better capturing interspeaker information by active learning for speaker identification,” in *IJCNN’01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, vol. 4. IEEE, 2001, pp. 2975–2980.
- [22] L. Wang, K. Chen, and H. Chi, “Capture interspeaker information with a neural network for speaker identification,” *IEEE Transactions on neural networks*, vol. 13, no. 2, pp. 436–445, 2002.
- [23] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, “A review of speaker diarization: Recent advances with deep learning,” *Computer Speech & Language*, vol. 72, p. 101317, 2022.
- [24] P.-A. Broux, D. Doukhan, S. Petitrenaud, S. Meignier, and J. Carrière, “An active learning method for speaker identity annotation in audio recordings,” in *1st International Workshop on Multimodal Media Data Analytics (MMDA 2016)*, 2016.
- [25] C. Yu and J. H. Hansen, “Active learning based constrained clustering for speaker diarization,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2188–2198, 2017.
- [26] S. Karlos, K. Kaleris, N. Fazakis, V. G. Kanas, and S. Kotsiantis, “Optimized active learning strategy for audiovisual speaker recognition,” in *International Conference on Speech and Computer*. Springer, 2018, pp. 281–290.
- [27] S. H. Shum, N. Dehak, and J. R. Glass, “Limited labels for unlimited data: Active learning for speaker recognition,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [28] M. Rouvier, R. Dufour, and P.-M. Bousquet, “Review of different robust x-vector extractors for speaker verification,” in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1–5.
- [29] T. Sainburg, L. McInnes, and T. Q. Gentner, “Parametric umap embeddings for representation and semi-supervised learning,” *arXiv preprint arXiv:2009.12981*, 2020.
- [30] G. A. Carpenter, S. Grossberg, and D. B. Rosen, “Fuzzy art: Fast stable learning and categorization of analog patterns by an adaptive resonance system,” *Neural networks*, vol. 4, no. 6, pp. 759–771, 1991.
- [31] A. Der Kiureghian and O. Ditlevsen, “Aleatory or epistemic? does it matter?” *Structural safety*, vol. 31, no. 2, pp. 105–112, 2009.
- [32] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” *arXiv preprint arXiv:1706.08612*, 2017.
- [33] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” *arXiv preprint arXiv:1806.05622*, 2018.
- [34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The kaldı speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.
- [35] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” *arXiv preprint arXiv:1510.08484*, 2015.
- [36] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký, “Building and evaluation of a real room impulse response dataset,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.