Closed-Set Speaker Identification using Few-Shot Transductive Learning

Gabriel Pîrlogeanu, Ana Neacșu, Horia Cucu
SpeeD, University Politehnica of Bucharest
Bucharest, Romania
{gabriel.pirlogeanu, ana_antonia.neacsu, horia.cucu}
@upb.ro

Jean-Christophe Pesquet
CVN, CentraleSupélec
Université Paris-Saclay, Inria
Gif-sur-Yvette, France
jean-christophe.pesquet@centralesupelec.fr

Ismail Ben Ayed
ÉTS Montréal
Montréal, Canada
Ismail.BenAyed@etsmtl.ca

***Abstract*—Closed-set unseen speaker identification is a crucial task in various applications, including forensics, fraud prevention, multi-speaker meetings, and speaker retrieval. The objective is to identify a specific individual at test time from a predefined group of speakers that were not seen during training, referred to as the watchlist. This task inherently aligns with the Few-Shot Learning paradigm. To address this challenge, we propose a novel few-shot transductive learning method, Few-Shot for A Single Class (FSAiC), which leverages a maximum likelihood approach specifically tailored for speaker identification. Furthermore, this study presents the first comprehensive evaluation of Few-Shot Speaker Identification, examining both inductive and transductive methods across both in-domain and out-of-domain scenarios. A key advantage of the proposed approach is its capacity to accommodate an arbitrary number of speakers in the watchlist while consistently outperforming state-of-the-art inductive and transductive algorithms, particularly in multi-shot settings.**

***Index Terms*—Speaker Identification, Closed-Set, Transductive Learning, Forensics**

I. INTRODUCTION

Speaker recognition, including speaker verification (SV), speaker identification (SI), and speaker diarization (SD), has been studied for over three decades [1]. Even though SV has dominated due to its role in biometric authentication [2], SI supports diverse applications, such as multi-speaker meeting identification [3], user-based customization for voice assistants [4], speaker naming in media [5]–[7], speaker retrieval [8]–[10], emergency call centers [11], and forensic or fraud prevention [12], [13].

While SI is a multiclass classification problem, real-world scenarios rarely have a fixed set of speakers, as new ones are continually added. Thus, identifying unseen speakers is more realistic. Unseen SI has two categories: closed-set and open-set. In the closed-set case, all utterances come from enrolled speakers (*i.e.*, in a watchlist) who were **not encountered during training**. Open-set Speaker Identification (OSSI) extends this by allowing utterances from speakers outside the watchlist, making it an even more challenging problem.

Recent research has focused on the OSSI task [14]–[17]. Shon *et al.* introduced the MCE2018 challenge [18] for OSSI, evaluating systems with top-S and top-1 metrics. The top submission [19] reported a 40% relative increase in top-1 error due to identification errors, a trend also noted in [16]. While most OSSI studies treat the task as two-step, they focus on verification and assume identification is trivial [14], despite previous studies suggesting otherwise. Thus, these studies may overlook practical challenges by focusing only on VoxCeleb distributions or small household scenarios (4-6 speakers).

However, closed-Set Unseen SI has also gained attention with the rise of meta-learning techniques for SR [20], with Few-Shot Prototypical Networks (PNs) being widely used [21]. However, Laenen *et al.* [22] showed that PNs can be outperformed by non-episodic inductive and transductive baselines [23], [24]. While inductive methods classify queries individually, transductive methods utilize the entire query set statistics [25], [26], leading to notable performance gains. These transductive methods often leverage pre-trained foundation models instead of complex episodic learning pipelines.

Therefore, we will focus only on closed-set unseen SI, to push benchmarks towards more realistic setups. The task involves identifying a speaker from a watchlist of enrolled speakers (K classes in the support set $\mathbb{S}$) based on one or more short utterances at test time (query set $\mathbb{Q}$). This scenario is inherently a Few-Shot Learning problem, with the key distinction that $\mathbb{Q}$ contains samples from a **single class**, becoming a **Closed-Set Transductive Few-Shot Single Speaker Identification** task. We assume that test-time utterances originate from a single speaker for the following practical reasons: (*i*) multi-channel setups, such as telephonic conversations, inherently manage speaker separation; (*ii*) in mono-channel multi-speaker scenarios, speaker diarization is often applied as preprocessing; and (*iii*) restricting test-time utterances to a single speaker enables a more robust transductive approach, eliminating the need for query distribution estimation when aggregating utterances.

We evaluate both inductive and transductive methods across watchlists of varying sizes in both in-domain and out-of-domain settings. Unlike traditional few-shot setups that restrict the support set $\mathbb{S}$ to a limited number of classes (e.g., 5- or 10-way classification), our experiments involve 500 to 1100 classes. To reflect the single-speaker case, we constrain $\mathbb{Q}$ to one effective class, with 1 to 5 short utterances per speaker. This setting necessitates a departure from existing few-shot approaches, which we address by proposing a novel, tuning-

free, optimization method that scales well to large support sets.

To summarize, our contributions are as follows: (*i*) we introduce a more realistic benchmark for closed-set unseen SI, (*ii*) we perform the first comprehensive evaluation of inductive and transductive methods for Few-Shot Closed-Set SI, and (*iii*) we propose a novel few-shot learning method that adapts better to the scenario where there is a single speaker or class in the query set. Note that all the code will be made available on GitHub[1] upon the acceptance of the paper.

The rest of the paper is structured as follows. The theoretical background of the methods explored in our work and the proposed methodological approaches are presented in Section II. In Section III-A, we introduce the datasets, training, and few-shot experimental setup. Lastly, in Section III-B we present our results, followed by a short conclusion in Section IV.

II. METHODS

A. Few-shot approach

Consider a labeled training set $\mathcal{D}_{\text{base}}$, defined as $\mathcal{D}_{\text{base}} = \{(\overline{\mathbf{x}}_n, \overline{y}_n)\}_{1 \leqslant n \leqslant |\mathcal{D}_{\text{base}}|}$, where $(\overline{\mathbf{x}}_n)_{1 \leqslant n \leqslant |\mathcal{D}_{\text{base}}|}$ represents the raw embeddings in $\mathbb{R}^d$ extracted using a feature extractor f_θ (parameterized by some vector θ), and $(\overline{y}_n)_{1 \leqslant n \leqslant |\mathcal{D}_{\text{base}}|}$ are the corresponding labels. In the few-shot learning literature, this labeled set is commonly referred to as the meta-training or base dataset. Let $\mathcal{Y}_{\text{base}}$ denote the set of classes within this base dataset. The few-shot setting assumes that a *test* dataset $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ is provided, using a notation similar to $\mathcal{D}_{\text{base}}$. However, $\mathcal{D}_{\text{test}}$ contains an entirely different set of classes $\mathcal{Y}_{\text{test}}$ such that $\mathcal{Y}_{\text{base}} \cap \mathcal{Y}_{\text{test}} = \varnothing$. From this test dataset, few-shot tasks are created by sampling a small number of labeled examples.

Specifically, each K-way $N_{\mathbb{S}}$-shot task involves sampling $N_{\mathbb{S}}$ labeled examples from each of K distinct classes, usually chosen at random. $\mathbb{S}$ denotes the index set of these labeled examples, known as the support set, with a total size $|\mathbb{S}| = K\, N_{\mathbb{S}}$. Additionally, each task includes a query set, $\mathbb{Q}$, indexing $N_{\mathbb{Q}}$ unlabeled (unseen) examples. In the context considered in this paper, the query set $\mathbb{Q}$ contains a single class (i.e., speaker). This differs from standard few-shot settings where $\mathbb{Q}$ may contain examples from the K classes.

Feature Normalization: Similar to Wang *et al.* [23], we L_2-normalize the support embeddings $(\mathbf{x}_n)_{n \in \mathbb{S}}$ and query embeddings $(\mathbf{x}_n)_{n \in \mathbb{Q}}$ produced by feature extractor f_θ.

B. Few-shot for a single class

We introduce a new tuning-free method called "**F**ew-**S**hot for **A** s**i**ngle **C**lass" (`FSAiC`), tailored to the assumption that the effective number of classes K_{eff} in the query set is equal to 1. For each sample $n \in \mathbb{S} \cup \mathbb{Q}$, we define the one-hot encoding vector $\mathbf{u}_n$ with components

$$(\forall k \in \{1, \ldots, K\}) \quad u_{n,k} = \begin{cases} 1 & \text{if } n \text{ is in class } k \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

[1]https://github.com/gabitza-tech/few-shot-si

We further assume the following Gaussian probabilistic model:

$$\mathbf{x}_n \sim \mathcal{N}(\mathbf{w}_k, \sigma^2 \boldsymbol{I}_d), \quad \text{if } u_{n,k} = 1, \tag{2}$$

where $\mathbf{w}_k \in \mathbb{R}^d$, $\sigma > 0$, and $\boldsymbol{I}_d$ is the identity matrix of size $d \times d$. Our objective is to find the class q of the data in the query set as well as estimating the means $(\mathbf{w}_k)_{1 \leqslant k \leqslant K}$. To do so, we adopt the maximum likelihood approach where we maximize the probability distribution of the feature vectors expressed as

$$\prod_{n \in \mathbb{S}} \Big(\sum_{k=1}^{K} u_{n,k}\, g(\mathbf{x}_n - \mathbf{w}_k \mid \sigma) \Big) \prod_{n \in \mathbb{Q}} g(\mathbf{x}_n - \mathbf{w}_q \mid \sigma),$$

where $g(\cdot \mid \sigma)$ denotes the probability density function of a zero-mean Gaussian vector with uncorrelated components of variance σ^2. This results in a mixed discrete-continuous optimization problem.

Moreover, we impose the condition that, for every $k \in \{1, \ldots, K\}$, the norm of $\mathbf{w}_k$ is equal to 1. The constrained maximum likelihood problem can then be addressed using the Lagrange multiplier method, yielding an explicit solution. To this end, we define the normalized centroid of the support class $k \in \{1, \ldots, K\}$ as $\mathbf{w}_k^{\mathbb{S}}$, and the centroid of the same class, under the assumption that the query samples belong to it, as $\mathbf{w}_k^{\mathbb{S} \cup \mathbb{Q}}$:

$$\mathbf{w}_k^{\mathbb{S}} = \frac{\sum_{n \in \mathbb{S}} u_{n,k} \mathbf{x}_n}{\| \sum_{n \in \mathbb{S}} u_{n,k} \mathbf{x}_n \|} \tag{3}$$

$$\mathbf{w}_k^{\mathbb{S} \cup \mathbb{Q}} = \frac{\sum_{n \in \mathbb{S}} u_{n,k} \mathbf{x}_n + \sum_{n \in \mathbb{Q}} \mathbf{x}_n}{\| \sum_{n \in \mathbb{S}} u_{n,k} \mathbf{x}_n + \sum_{n \in \mathbb{Q}} \mathbf{x}_n \|}. \tag{4}$$

After some simple calculations, it can be shown that the optimal solution to the constrained maximum likelihood problem is obtained by finding $q \in \{1, \ldots, K\}$ minimizing

$$C_q = \sum_{n \in \mathbb{S}} u_{n,q} (\|w_q^{\mathbb{S} \cup \mathbb{Q}} - \mathbf{x}_n\|^2 - \|w_q^{\mathbb{S}} - \mathbf{x}_n\|^2) + \sum_{n \in \mathbb{Q}} \|w_q^{\mathbb{S} \cup \mathbb{Q}} - \mathbf{x}_n\|^2, \tag{5}$$

while the associated mean estimates are $\mathbf{w}_k = \mathbf{w}_k^{\mathbb{S}}$ if $k \neq q$ and $\mathbf{w}_q = \mathbf{w}_q^{\mathbb{S} \cup \mathbb{Q}}$. The optimal q value is found by minimizing C_q. This method provides a streamlined approach, balancing the complexity of implementation with the need for effective class probability estimation in few-shot learning scenarios.

C. Other methods

For a comprehensive evaluation, we compared our approach with the established inductive baseline method `SimpleShot` [23] and several variations we introduced to address our specific scenario. Additionally, we assessed our method against a majority vote version of `PADDLE` [27], detailed below. We do not compare our methods with PNs, as they are very sensitive to the $\mathbb{Q}$-$\mathbb{S}$ training episode configuration and appear unsuitable for our scenario, where $K_{\text{eff}} \ll K$.

- **SimpleShot** (`SS`): Based on the nearest neighbor rule, a class q is assigned to each query sample $\mathbf{x}_n$ with $n \in$

$\mathbb{Q}$ based on the smallest Euclidean distance to $\mathbf{w}_q^{\mathbb{S}}$ the support class q centroid from (3). For $(\mathbf{x}_n)_{n\in\mathbb{Q}}$, we assign labels $(q_n)_{n\in\mathbb{Q}}$ in the query, following the assignment rule:

$$(\forall n \in \mathbb{Q}) \quad q_n = \underset{k\in\{1,\dots,K\}}{\arg\min} \|\mathbf{x}_n - \mathbf{w}_k^{\mathbb{S}}\|. \tag{6}$$

This method is used as an inductive baseline because it provides an estimated label for each query sample independently from other samples in the query set $\mathbb{Q}$.

- **SimpleShot majority vote** (SMV): Exploiting the fact that $\mathbb{Q}$ contains only one class, we implement a majority voting variation of SimpleShot that assigns a single speaker to all query samples $(\mathbf{x}_n)_{n\in\mathbb{Q}}$.
- **PADDLE:** We also evaluate our method against a variation of a state-of-the-art transductive iterative approach, where we incorporate majority voting. Martin *et. al.* [27] recently introduced the **P**rim**A**l **D**ual **M**inimum **D**escription **LE**ngth (PADDLE) algorithm to address the realistic scenario in few-shot learning where the number of effective classes K_{eff} in the query set is significantly smaller than in the support set, which is indeed our context. The authors assumed a Gaussian Mixture Model for the query set data, leading to a statistical approach based on the Expectation Maximization algorithm. Their method effectively tackles a general Minimum Description Length problem balancing data-fitting accuracy and model complexity. To avoid parameter tuning, the penalization parameter has been set to its theoretical value, $\lambda = N_{\mathbb{Q}}$.

III. SIMULATION RESULTS

A. Experimental setting

Datasets: We use several large-scale datasets for both training and evaluation, in order to conduct a comprehensive analysis. For training, we use the dev set from VoxCeleb2 [28], with 5994 speakers and 1M utterances of variable lengths, as well as the chinese multi-genre dataset CN-Celeb2 [29], with 1996 speakers and approximately 500k utterances. We also use the RIR [30] and MUSAN [31] datasets for data augmentation. We evaluate across several in-domain and out-of-domain datasets in order to accurately represent the performance of different inductive and transductive methods.

VoxCeleb1 [32]: Containing 1251 speakers, approximately 150k utterances. It has a similar data distribution to the VoxCeleb2 dataset. We divided the dataset classes in two subsets: 10% development classes, for model fine-tuning or hyper-parameter tuning, and 90% test classes. This results in 1125 speakers in the test set.

CN-Celeb1 [33]: Containing a total of 997 speakers and about 130k recordings. Out of these 997 speakers, we only keep the speakers that have at least 10 samples, meaning we are left with 951 speakers and approximately 126k recordings. Similarly, we apply a 10%-90% dev-test split, resulting in 856 speakers in the test set.

JukeBox-V1 [34]: Containing a total of 670 singers and a total of 385 hours of data, obtained from the combination of the auxiliary, train and test subsets. Out of these 670 classes, we only keep the classes with at least 10 samples, resulting in 561 classes and approximately 43k recordings. We apply a 10%-90% dev-test split, resulting in 505 speakers in the test set. We conduct this singer recognition evaluation as it represents a challenging out-of-domain scenario, with voice variations from both pitch changes and background sounds.

TABLE I
TOP 1 ACCURACY (%) RESULTS ON THE VOXCELEB1 DATASET.
BOLD = BEST RESULT; UNDERLINE = SECOND BEST RESULT.

$N_{\mathbb{Q}}$	SS	SMV	PADDLE	FSAiC
1	92.38	92.38	<u>92.42</u>	**92.48**
3	91.7	96.81	<u>99.06</u>	**99.18**
5	91.2	98.45	<u>99.41</u>	**99.56**

Each recording in the evaluation datasets is randomly cropped to a 3s audio, with the exception of the JukeBox-V1 dataset, where we split the recordings in 30s non-overlapping segments, similar to the original paper.

Model training: We have chosen to use an ECAPA-TDNN [35] architecture with filter length $C = 1024$, totalling 14.7M parameters, as our feature extractor. We train two models from scratch, one on 5994 speakers from VoxCeleb2 dev set, and another on 7990 speakers from VoxCeleb2 dev and CN-Celeb2 dataset. The recordings used for training are split in non-overlapping segments of 3s, resulting in 2.27M utterances for the first model, and 3.35M utterances for the second model, respectively. We follow a similar training setup to [36], where we apply online augmentation using simulated RIR and MUSAN noise, as well as SpecAugment [37]. We train each model with AAM Softmax, setting a margin of 0.2 and a scale of 30, for 25 epochs with a batch size of 1024, using an A100 80 GB GPU.

The model trained only on VoxCeleb2 achieves an Equal Error Rate (EER) of 1.21% on Vox-O trial set and an EER of 15.25% on CN-Celeb1 trial set. The model trained on both VoxCeleb2 and CN-Celeb2 achieves an EER of 1.2% on Vox-O and 8.07% on CN-Celeb1. We do not use any score normalisation.

Few-shot setup: For all the datasets analyzed, we extract embeddings from the recordings using a feature extractor f_θ as explained above. We construct our few-shot evaluation tasks with varying numbers of $N_{\mathbb{S}}$ shots and $N_{\mathbb{Q}}$ query samples, with values ranging from 1 to 5. Unlike standard few-shot evaluation setups, we do not constrain our episodes to 5- or 10-way classification. Instead, for all evaluation datasets, we utilize the entire test subset, which includes between 505-ways and 1125-ways classification. Following standard few-shot evaluation practices, we report the Top-1 accuracy averaged over 10,000 tasks.

B. Results

We first evaluate the simplest scenario on the VoxCeleb1 dataset, reporting 3-shot 1125-way results in Table I. Varying

TABLE II

RESULTS ON CN-CELEB1 AND JUKEBOX-V1 DATASETS. THE COLUMN *Train* INDICATES THE DATASETS ON WHICH THE FEATURE EXTRACTOR f_θ WAS TRAINED. $N_{\mathbb{S}}$ REPRESENTS THE NUMBER OF SHOTS IN THE SUPPORT SET $\mathbb{S}$ AND $N_{\mathbb{Q}}$ REPRESENTS THE NUMBER OF SAMPLES IN $\mathbb{Q}$. THE LAST TWO ROWS OF THE TABLE REPORT THE AVERAGE ACCURACY OF THE Vox2 AND Vox2+CN2 MODELS FOR EACH METHOD, ACROSS ALL $N_{\mathbb{S}}$:$N_{\mathbb{Q}}$ CONFIGURATIONS. **BOLD** = BEST RESULT; UNDERLINE = SECOND BEST RESULT. WE REPORT TOP 1 ACCURACY (%).

		CN-Celeb1				**JukeBox-V1**			
Train	$N_{\mathbb{S}}$:$N_{\mathbb{Q}}$	SS	SMV	PADDLE	FSAiC	SS	SMV	PADDLE	FSAiC
Vox2	1:1	**23.09**	**23.09**	**23.09**	**23.09**	**29.64**	**29.64**	**29.64**	**29.64**
	1:3	21.36	25.17	**35.85**	35.18	28.27	32.44	**42.24**	39.38
	1:5	20.58	31.03	**40.43**	39.75	27.49	37.50	**45.58**	42.02
	3:1	36.40	36.40	35.66	**36.44**	37.43	37.43	38.25	**38.34**
	3:3	34.84	40.93	51.10	**54.06**	36.01	41.49	50.87	**52.19**
	3:5	34.31	49.80	56.28	**60.62**	35.37	47.20	55.35	**57.54**
	5:1	42.17	42.17	41.92	**42.41**	41.10	41.10	41.78	**41.92**
	5:3	41.21	48.09	59.42	**61.10**	39.46	45.14	**57.49**	57.38
	5:5	40.50	57.41	65.19	**67.89**	38.85	51.33	62.22	**63.21**
Vox2+CN2	1:1	**34.35**	**34.35**	**34.35**	**34.35**	**34.46**	**34.46**	**34.46**	**34.46**
	1:3	31.24	36.64	**48.61**	47.92	32.30	36.69	**47.97**	46.77
	1:5	29.73	42.98	**52.26**	51.40	31.33	42.28	**50.57**	49.26
	3:1	49.19	49.19	48.63	**49.41**	46.08	46.08	46.89	**46.99**
	3:3	46.29	53.89	63.41	**66.33**	43.53	49.39	60.02	**61.22**
	3:5	44.72	61.69	67.31	**71.70**	42.59	56.05	63.62	**66.42**
	5:1	54.47	54.47	54.46	**54.89**	51.36	51.36	51.86	**51.87**
	5:3	51.68	59.45	69.52	**71.95**	47.86	54.18	65.76	**66.49**
	5:5	50.40	67.96	74.46	**77.65**	46.35	60.24	69.73	**71.42**
AVG Vox2		32.72	39.34	45.44	**46.63**	34.85	40.36	**47.05**	46.85
AVG Vox2+CN2		43.56	51.18	57.00	**58.40**	41.76	47.86	54.54	**54.96**

$N_{\mathbb{Q}}$ between 1 and 5 allows us to evaluate the influence of the available query set. The model was trained on VoxCeleb2, representing an in-domain scenario. The proposed method FSAiC consistently outperforms the other evaluated methods and manages to relatively improve by up to 9.1% over the inductive baseline. Notably, with a few labelled samples, both PADDLE and FSAiC reach around 99% Top-1 accuracy.

In Table II, we present results on the CN-Celeb1 dataset (856-ways) and the singing JukeBox-V1 dataset (505-ways). For both datasets, we evaluate multiple $\mathbb{Q}$-$\mathbb{S}$ configurations using two models. Compared to the VoxCeleb1 scenario, the 3-shot configuration exhibits a substantial performance drop—exceeding 30%—across all methods, models, and datasets. This confirms our hypothesis that prior OSSI benchmarks did not adequately address the most realistic closed-set subtask configurations.

Firstly, we observe a significant improvement across both datasets when using the model trained on English and Chinese. While expected for CN-Celeb1, which benefits from Chinese-language data, a similar trend on the singing dataset suggests that CN-Celeb2's diverse domain coverage, including singing, may contribute to this effect. Moreover, transductive approaches consistently outperform the inductive baseline SS, SSMV achieving an absolute Top-1 accuracy improvement of up to 17%.

However, PADDLE and FSAiC exhibit greater generalization capabilities even for out-of-domain tasks. In the 1-shot scenario, PADDLE achieves the best overall performance, even surpassing FSAiC, itself significantly outperforming the inductive baseline SS. The improvement is particularly notable in the most challenging experimental condition—the JukeBox-V1 dataset evaluated with the Vox2 model—wherein PADDLE demonstrates superior robustness when confronted with significant domain shifts.

Finally, in the multi-shot scenario, FSAiC consistently outperforms all other methods. Compared to the inductive baseline, it achieves an absolute accuracy improvement of up to 26.3% and compared to the transductive baseline PADDLE, it exhibits a gain of up to 4.4%. Averaging across all configurations and datasets, FSAiC achieves the best overall results. Therefore, FSAiC stands as an effective baseline for Single Query-Class Few-Shot tasks, offering strong out-of-domain robustness with reasonable complexity.

IV. CONCLUSIONS

This paper presents the first comprehensive study on large-scale closed-set Few-Shot Speaker Identification in both in-domain and out-of-domain scenarios, demonstrating the advantages of transductive methods over inductive ones. Our proposed method, FSAiC, consistently outperforms competing approaches in multi-shot scenarios and demonstrates superior robustness in out-of-domain conditions. Additionally, we explore a variation of the transductive method PADDLE, which performs very well in the one-shot setting. Despite these

significant advancements, our findings emphasize that Closed-Set SI remains a challenging research area. Future work will extend these investigations to open-set identification and in the wild multi-speaker mono channel conversations.

This work was partly funded by EU HORIZON project no. 101070190 (AI4TRUST project) and by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI - UEFISCDI, project number PN-IV-P8-8.1-PRE-HE-ORG-2023-0078, within PNCDI IV.

REFERENCES

[1] M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities," *IEEE Access*, vol. 9, pp. 79 236–79 263, May 2021.

[2] W. L. Y. Tu and M.-W. Mak, "A survey on text-dependent and text-independent speaker verification," *IEEE Access*, vol. 10, pp. 99 038–99 049, Sep. 2022.

[3] G. Biagetti, P. Crippa, L. Falaschetti, and S. Orcioni, "Robust speaker identification in a meeting with short audio segments," Tenerife, Spain, June 2016.

[4] R. Li, J.-Y. Jiang, X. Wu, C.-C. Hsieh, and A. Stolcke, "Speaker identification for household scenarios with self-attention and adversarial training," in *Interspeech*, Shanghai, China, 25-29 Oct. 2020.

[5] Y. Hu, J. S. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang, "Deep multimodal speaker naming," in *Proceedings of the 23rd ACM international conference on Multimedia*, Brisbane, Australia, Oct. 2015, p. 1107–1110.

[6] M. L. Bellagha and M. Zrigui, "Speaker naming in tv programs based on speaker role recognition," in *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, 2020, pp. 1–8.

[7] M. Azab, M. Wang, M. Smith, N. Kojima, J. Deng, and R. Mihalcea, "Speaker naming in movies," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, Jun. 2018, pp. 2206–2216.

[8] S. Shon, Y. Lee, and T. Kim, "Large-scale speaker retrieval on random speaker variability subspace," in *Interspeech*, Graz, Austria, 15-19 Sep. 2019, pp. 2963–2967.

[9] X. Su, Q. Zhan, C. Hu, and X. Xie, "Combination of multiple embeddings for speaker retrieval," in *The Speaker and Language Recognition Workshop (Odyssey 2022)*, Beijing, China, 28 June-1 July 2022, pp. 384–389.

[10] E. Loweimi, M. Qian, K. Knill, and M. Gales, "On the usefulness of speaker embeddings for speaker retrieval in the wild: A comparative study of x-vector and ecapa-tdnn models," in *Interspeech 2024*, Kos Islands, Greece, 1-5 Sep. 2024, pp. 3774–3778.

[11] J. Gałka, J. Równicka, M. Igras-Cybulska, P. Jaciów, K. Wajda, M. Witkowski, and M. Ziolko, "System supporting speaker identification in emergency call center," in *Interspeech 2024*, Dresden, Germany, 6-10 Sep. 2015.

[12] A. A. Moura, N. Nepomuceno, and V. Furtado, "Enhancing speaker identification in criminal investigations through clusterization and rank-based scoring," *Forensic Science International: Digital Investigation*, 9-12 July 2024.

[13] G. S. Morrison, F. H. Sahito, G. Jardine, D. Djokic, S. Clavet, S. Berghs, and C. Goemans Dorny, "Interpol survey of the use of speaker identification by law enforcement agencies," *Forensic Science International*, Mar. 2016.

[14] R. Peri, S. O. Sadjadi, and D. Garcia-Romero, "Voxwatch: An open-set speaker recognition benchmark on voxceleb," June 2023. [Online]. Available: https://arxiv.org/abs/2307.00169

[15] K. C. Kishan, Z. Tan, L. Chen, M. Jin, E. Han, A. Stolcke, and C. Lee, "OpenFEAT: Improving speaker identification by open-set few-shot embedding adaptation with transformer," in *IEEE Int. Conf. Acoust., Speech, Signal. Process*, Singapore, Australia, 22-27 Feb. 2022, pp. 7062–7066.

[16] K. Wilkinghoff, "On open-set speaker identification with I-vectors," in *The Speaker and Language Recognition Workshop*, Tokyo, Japan, 1-5 Nov. 2020, pp. 408–414.

[17] Y. Lin, M. Cheng, F. Zhang, Y. Gao, S. Zhang, and M. Li, "Voxblink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark," in *Interspeech 2024*, 2024, pp. 4263–4267.

[18] S. Shon, N. Dehak, D. Reynolds, and J. Glass, "MCE 2018: The 1st multi-target speaker detection and identification challenge evaluation," in *Interspeech*, Graz, Austria, 15-19 Sep. 2019, pp. 356–360.

[19] E. Khoury, K. Lakhdhar, A. Vaughan, G. Sivaraman, and P. Nagarsheth, "Pindrop Labs' submission to the first multi-target speaker detection and identification challenge," in *Interspeech*, Graz, Austria, 15-19 Sep. 2019, pp. 1502–1505.

[20] J. S. Chung, J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Interspeech*, Shanghai, China, 25-29 Oct. 2020, pp. 2977–2981.

[21] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Red Hook, NY, USA, 4-9 Dec. 2017, p. 4080–4090.

[22] S. Laenen and L. Bertinetto, "On episodes, prototypical networks, and few-shot learning," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21, Red Hook, NY, USA, 6-14 Dec. 2021.

[23] Y. Wang, W.-L. Chao, K. Q. Weinberger, and L. V. D. Maaten, "SimpleShot: Revisiting nearest-neighbor classification for few-shot learning," *preprint arXiv:1911.04623*, Aug. 2019.

[24] Y. Chen, Z. Liu, H. Xu, T. Darrell, and X. Wang, "Meta-baseline: Exploring simple meta-learning for few-shot learning," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, 10-17 Oct. 2021, pp. 9042–9051.

[25] O. Veilleux, M. Boudiaf, P. Piantanida, and I. B. Ayed, "Realistic evaluation of transductive few-shot learning," *ArXiv*, vol. abs/2204.11181, Apr. 2022.

[26] L. Tian, J. Feng, X. Chai, W. Chen, L. Wang, X. Liu, and B. Chen, "Prototypes-oriented transductive few-shot learning with conditional transport," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Vancouver, Canada, 18-22 June 2023, pp. 16 317–16 326.

[27] S. Martin, M. Boudiaf, E. Chouzenoux, J.-C. Pesquet, and I. B. Ayed, "Towards practical few-shot query sets: Transductive minimum description length inference," in *Adv. Neur. Inform. Proc. Syst.*, vol. 35, New Orleans, USA, 28 Nov.-9 Dec. 2022, pp. 34 677–34 688.

[28] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Interspeech*, Hyderabad, India, 2-6 Sep. 2018, pp. 1086–1090.

[29] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "CN-Celeb: Multi-genre speaker recognition," *Speech Comm.*, vol. 137, pp. 77–91, Feb. 2022.

[30] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE Int. Conf. Acoust., Speech, Signal. Process*, New Orleans, USA, 5-9 March 2017, pp. 5220–5224.

[31] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *preprint arXiv:1510.08484*, Oct. 2015.

[32] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Interspeech*, Stockholm, Sweden, 20-24 Aug. 2017, pp. 2616–2620.

[33] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "CN-Celeb: A challenging chinese speaker recognition dataset," in *IEEE Int. Conf. Acoust., Speech, Signal. Process*, Barcelona, Spain, 4-8 May 2020, pp. 7604–7608.

[34] A. Chowdhury, A. Cozzo, and A. Ross, "JukeBox: A multilingual singer recognition dataset," in *Interspeech*, Shanghai, China, 25-29 Oct. 2020, pp. 2267–2271.

[35] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, Shanghai, China, 25-29 Oct. 2020, pp. 3830–3834.

[36] R. K. Das, R. Tao, and H. Li, "HLT-NUS submission for 2020 NIST conversational telephone speech SRE," *preprint arXiv:2111.06671*, Nov. 2021.

[37] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, Graz, Austria, 15-19 Sep. 2019, pp. 2613–2617.