



Towards Secure User Authentication for Headphones via In-Ear or In-Earcup Microphones

N Shashaank^{1,2}, *Xiao Quan*¹, *Andrew Kaluzny*¹, *Leonard Varghese*¹, *Marko Stamenovic*¹,
*Chuan-Che (Jeff) Huang*¹

¹Bose Corporation, Boston, MA, USA

²Department of Computer Science, Columbia University, New York, NY, USA

shashaank.n@columbia.edu,

{xiao-quan, andrew.kaluzny, leonard.varghese, marko.stamenovic, chuan-che.huang}@bose.com

Abstract

As headphones and earbuds with integrated AI assistants become more prevalent, it is important to prevent other people from gaining unauthorized access to these devices and accessing personal, sensitive user information. One solution to this problem is to implement a speaker identification (SID) model that can register authorized users onto the headphones and verify their identity in real-time, but it is challenging for these models to work in loud and/or noisy conditions, and they can be easily hacked using voice cloning, spoofing, and other adversarial attack techniques. In this paper, we propose using speech data collected from in-ear and in-earcup microphones commonly found on noise-cancelling headphones and earbuds to fine-tune SID models to address these issues. We collected inside mic data from 195 speakers across 4 headphones and earbuds and show that fine-tuning the model on multiple devices can improve performance when evaluating on a single device.

Index Terms: user authentication, speaker identification, in-earcup microphones, neural networks, few-shot learning

1. Introduction

Recent advances in the capabilities of large language models and AI voice assistants are making it possible for people to perform simple tasks that require accessing private information on their phones and mobile devices using just voice commands, such as ordering a meal or checking their bank information. While these models are currently too large to run on-device on audio wearable devices such as headphones or earbuds, self-voice detection systems [1, 2] allow users to interact with these AI assistants through their headphones using voice commands and speech prompts, accessing information through paired mobile devices or directly through Wi-Fi without an intermediate device. However, if other people gain access to users' headphones or earbuds (e.g. by accidentally losing them or leaving them unattended for an extended period of time), they could access the same personal information without the proper authorization, as self-voice detection only determines whether the current speaker is directly wearing the headphones or not and cannot differentiate between multiple speakers.

A potential solution to this problem is to deploy a speaker identification (SID) system that can accept authorized users and reject unknown users based on their vocal characteristics. Typically, these systems have an *enrollment* phase to register users by recording a small sample of their speech and converting them into speaker "embeddings" and a *verification* phase that runs in real-time to authenticate the user's identity based on the speaker embeddings and grant access to the device. Recent work has shown training neural networks for SID achieves great performance [3, 4], and it is possible to create lightweight neural

networks [5, 6] that can run on low-power embedded devices with limited computing power typically found on headphones and earbuds while still achieving comparable performance to the state-of-the-art models. However, it is challenging for these models to work consistently in noisy and/or loud environments, and they are vulnerable to voice cloning, spoofing, and other adversarial attacks meant to fool the model into believing that an authenticated user is trying to access the device [7, 8].

In contrast to traditional "outside" microphones that capture sound propagated through airwaves, microphones located on the inside of the earcup or earbud (which we will refer to as "inside" mics) capture sound propagated through bone conduction and articulatory organs such as the vocal cords and vocal tract [2, 9]. Inside mics are acoustically shielded from outside noise and loud talkers, creating a moderate-to-high signal-to-noise ratio (SNR) with respect to the headphone wearer's self-speech. Depending on the specific setup, audio captured by inside mics can have very different spectral and temporal characteristics from outside mics, such as very little high-frequency information and quick attenuation of sounds from a short distance (Figure 1), which makes it difficult to apply voice cloning or spoofing to break the system. Currently, there has been little research into whether existing SID models will perform well on inside mics that have different characteristics from the original training data. Using inside mics also poses new questions around cross-device generalization, such as whether a model tailored for an in-ear form factor will generalize to an over-the-ear form factor, and whether deploying a model to a new form factor requires collecting data from that specific device or reusing data collected from other existing form factors.

In this paper, we present three contributions to improve user authentication on headphones and earbuds using inside microphones. First, we characterize the performance of a traditional SID model on speech data recorded through inside mics and show that it degrades significantly compared to speech recorded from outside mics. Second, we fine-tune the model on inside mic speech data and show that it outperforms the baseline by up to 50 – 81% depending on the specific metric. Finally, we collect data from 195 speakers across 4 headphones and earbuds and show that fine-tuning the model on multiple form factors improves performance compared to fine-tuning on a single form factor, conditional on the target form factor being included in training. We also show that fine-tuning on over-the-ear headphones helps improve performance even on in-ear earbuds.

2. Methods

2.1. Prototypical Network

We selected the prototypical network [10] for our SID model, as it can accurately classify samples using limited input and

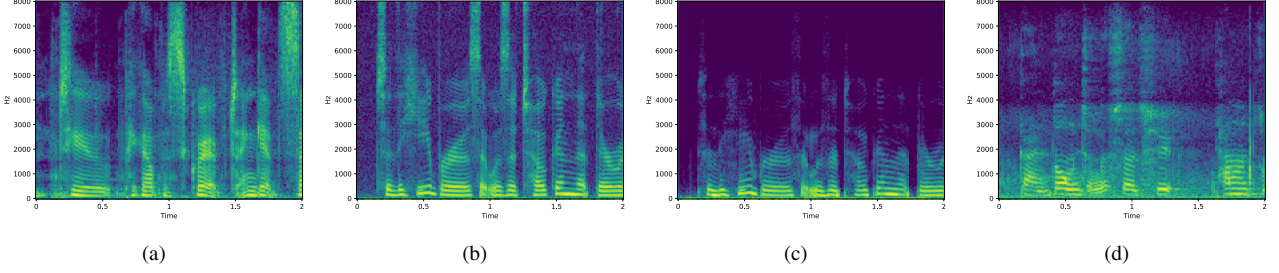


Figure 1: *Short-time Fourier transform spectrograms of speech recorded from outside microphones and inside microphones. (a) Standard speech example from the VoxCeleb2 dataset. (b) Speech example recorded through the outside mic from in-ear wireless earbuds in the BWAD dataset. (c) Speech example recorded through the inside mic from in-ear wireless earbuds in the BWAD dataset. (d) Speech example recorded through the inside mic from the Sony WF-1000XM3 Earbuds in the EarSAVAS dataset.*

adapt to unknown classes during inference. A recent literature review of models and loss functions for SID [11] showed that the prototypical network had the best performance compared to other models, and it has been adopted in several works for sound event detection and speaker identification [12, 13, 14].

Prototypical networks are trained on “episodes,” a type of minibatch commonly used for few-shot learning tasks. Each episode contains a randomly selected subset of speakers K of size C from the dataset, a support set $S = \{(\mathbf{x}_s, y_s)\}_{s=1}^{NC}$ with N random samples for each speaker, and a query set $Q = \{(\mathbf{x}_q, y_q)\}_{q=1}^{MC}$ with M random samples for each speaker, where \mathbf{x} is the input audio example and $y \in K$ is the speaker label. The support and query set samples are transformed into feature embeddings in a D -dimensional latent space using a neural network encoder E_θ with learnable parameters θ .

For each speaker $k \in K$, a “prototype” or centroid $\mathbf{c}_k \in \mathbb{R}^D$ is created from the speaker’s associated samples in the support set $S_k \subseteq S$ by computing the mean of the feature embeddings generated by the encoder, effectively describing the average representation of the speaker in the latent space:

$$\mathbf{c}_k = \frac{1}{N} \sum_{(\mathbf{x}_s, y_s) \in S_k} E_\theta(\mathbf{x}_s) \quad (1)$$

During inference, the feature embeddings from the query set are compared to the speaker prototypes using a distance function d , the result of which is transformed into probabilities using a softmax distribution. The speaker label with the closest prototype (and consequently the highest probability in the softmax) is then assigned as the class prediction.

$$p(y = k | \mathbf{x}) = \frac{\exp(d(E_\theta(\mathbf{x}), \mathbf{c}_k))}{\sum_{k'} \exp(d(E_\theta(\mathbf{x}), \mathbf{c}_{k'}))} \quad (2)$$

The loss function \mathcal{L} for each episode is computed as the cross entropy loss over the posterior probabilities from the softmax distribution, with the goal being to update θ for the encoder to learn the optimal latent space to generate feature embeddings:

$$\mathcal{L} = -\frac{1}{MC} \sum_{(\mathbf{x}_q, y_q) \in Q} \log p(y = y_q | \mathbf{x}_q) \quad (3)$$

We selected the ECAPA-TDNNLite model [5] as the architecture for E_θ , as it has been shown to achieve state-of-the-art performance on traditional speaker identification tasks while being lightweight enough to run on low-power embedded devices. We set $D = 256$ and modified the last layer of the model to return feature embeddings of the same size. We selected the

angular distance function [11] for d , a cosine similarity function with learnable scale and bias parameters.

2.2. Bose Wearable Audio Dataset

Our primary source of in-ear and in-earcup mic data used to develop and fine-tune our model is the internal Bose Wearable Audio Dataset (BWAD), which includes multi-channel recordings collected in an anechoic chamber from 4 Bose headphones and earbuds: 2 over-the-ear headphones, and 2 in-ear wireless earbuds. All data was gathered with active noise cancellation disabled, and it is part of a larger effort to collect relevant data from headphones through internal user studies.

We limited the selected recordings to continuous speech, and overlapping speakers between form factors were removed to avoid sampling the same speaker twice in a single data batch. In total, we had 195 speakers and 32 hours of audio, with 94 speakers from over-the-ear headphones and 101 speakers from in-ear wireless earbuds. As most recordings in the dataset were relatively long (> 30 s), we split files longer than 10 seconds into smaller segments ranging from 2 – 4 seconds to have enough samples to select during training and evaluation.

2.3. EarSAVAS Dataset

We also used in-ear mic data from the EarSAVAS dataset [15], an open-source dataset with recordings from modified Sony WF-1000XM3 Earbuds collected in a standard conference room with some ambient noise. The frequency response of recordings in this dataset (Figure 1d) are considerably different from the recordings in BWAD (Figure 1c), as well as the recordings from the bone-conduction mic described in EarVoice [2]. We again limited the selected recordings to continuous speech, and in total we had 43 speakers and 1.94 hours of audio. Similar to the process with BWAD, files longer than 10 seconds were split into smaller segments.

3. Experimental Setup

3.1. Dataset and Input Processing

In addition to BWAD and EarSAVAS, we also included the VoxCeleb2 [17] and VoxCeleb1 [16] datasets for training and evaluation on speech recorded using outside microphones. For VoxCeleb2, we used only the dev split which contains 5,994 speakers and 2,360 hours of audio, while for VoxCeleb1 we used both the dev and test splits containing 1,251 speakers and 351.6 hours of audio. For BWAD and EarSAVAS, we created a

Table 1: Evaluation of model pre-training and fine-tuning configurations. For each test dataset, the EER is reported from the speaker pair evaluation task, the 2-way Acc/AUROC is reported from the few-shot evaluation task with 2 closed-set speakers, and the 3-way Acc/AUROC is reported from the few-shot evaluation task with 3 closed-set speakers.

Test Dataset	Metric	Model Training Configurations					
		VoxCeleb2 Pre-Training	BWAD Single Fine-Tuning	BWAD OE Fine-Tuning	BWAD IE Fine-Tuning	BWAD All Fine-Tuning	EarSAVAS Fine-Tuning
VoxCeleb1 [16]	EER	3.75%	35.9%	34.2%	37.1%	38.5%	31.8%
	2-way Acc/AUROC	99.5%/72.0%	76.5%/48.5%	75.2%/50.9%	75.4%/49.7%	74.8%/48.7%	82.1%/47.4%
	3-way Acc/AUROC	99.1%/55.5%	63.1%/50.1%	63.1%/50.8%	64.3%/49.0%	64.0%/48.4%	72.7%/43.4%
BWAD Single	EER	23.3%	17.7%	14.5%	17.4%	11.7%	25.1%
	2-way Acc/AUROC	95.7%/41.8%	95.7%/62.8%	97.4%/65.3%	97.1%/63.3%	97.5%/73.4%	91.1%/45.9%
	3-way Acc/AUROC	92.5%/40.0%	92.1%/49.1%	94.9%/48.1%	94.5%/49.3%	95.8%/56.5%	84.4%/43.0%
BWAD OE	EER	23.2%	17.9%	14.9%	17.2%	12.3%	24.2%
	2-way Acc/AUROC	95.8%/41.2%	95.2%/63.5%	97.3%/62.9%	97.2%/61.5%	97.7%/74.5%	90.8%/44.8%
	3-way Acc/AUROC	92.1%/41.2%	92.1%/49.2%	95.1%/48.7%	94.4%/49.5%	96.0%/58.5%	84.3%/42.3%
BWAD IE	EER	22.8%	17.7%	14.3%	17.2%	11.7%	24.7%
	2-way Acc/AUROC	95.2%/42.1%	95.6%/63.5%	97.7%/65.5%	97.0%/63.5%	97.5%/74.6%	91.4%/46.0%
	3-way Acc/AUROC	92.5%/40.3%	91.8%/49.5%	95.4%/49.4%	94.6%/49.5%	95.9%/57.5%	84.1%/42.9%
EarSAVAS [15]	EER	14.9%	29.6%	26%	30.6%	32.8%	14.5%
	2-way Acc/AUROC	97.6%/47.1%	86.1%/46%	85.9%/47.3%	85.6%/47.7%	86.4%/46.1%	96.0%/82.9%
	3-way Acc/AUROC	95.5%/36.7%	78.9%/44.9%	76.2%/46.5%	76%/47.3%	76.6%/46.9%	92.6%/69.5%

randomized train/test split for each headphone/earbud form factor available within the dataset, with 15 speakers in the test split and the remaining speakers in the train split.

Audio files were standardized to have a 16 kHz sampling rate and downsampled where applicable to meet this standard. For BWAD and EarSAVAS, multi-channel audio files were converted to mono by extracting the data specifically recorded through the in-ear or in-earcup mic. For a given audio file, we randomly sampled a 2 s segment, converted it into an 80-bin log Mel spectrogram with a 25 ms window size and 10 ms hop size, and applied 1-dimensional instance normalization over the frequency dimension before passing it as input to the model.

3.2. Model Training

To establish a baseline model for SID without inside microphones, we first pre-trained the model on VoxCeleb2 for a total of 1,000,000 episodes. Each episode contained a random subset of $C = 200$ speakers, $N = 1$ support sample, and $M = 1$ query sample per speaker. We used the AdamW optimizer [18] for stochastic gradient descent, with an initial learning rate of 0.001 and a decay rate of 95% every 20,000 episodes.

Building on the pre-trained model, we then applied fine-tuning using either BWAD or EarSAVAS for a total of 20,000 episodes, 2% of the episodes used for pre-training. In contrast to pre-training, each episode had a random subset of only $C = 3$ speakers, $N = 5$ support samples, and $M = 5$ query samples per speaker. We chose 3 speakers per episode as a realistic upper bound on the number of speakers that a user might register for authentication on a single pair of headphones, and we selected 5 support samples to generate the prototype for each speaker as we expect users to record multiple examples/prompts for the model. We again used the AdamW optimizer but with no learning rate decay.

To understand the performance of fine-tuning with a single or multiple headphone/earbud form factors, we created 5 configurations based on our inside mic datasets. *BWAD Single* uses data from a single in-ear wireless earbud form factor in BWAD, *BWAD OE* uses data from the 2 over-the-ear headphones in BWAD, *BWAD IE* uses data from the 2 in-ear wireless earbuds in BWAD, *BWAD All* uses data from all 4 devices

in BWAD, and *EarSAVAS* uses data only from the single Sony WF-1000XM3 Earbuds in EarSAVAS.

3.3. Evaluation Tasks and Metrics

To quantify the performance of our model pre-training and fine-tuning configurations, we created two evaluation tasks: few-shot evaluation and speaker pair evaluation.

The *few-shot evaluation* task assesses the model’s performance in differentiating multiple registered speakers from each other and rejecting unknown speakers. It consists of 1,000 episodes, with two separate variations allotting 2 or 3 closed-set (i.e. registered) speakers and 10 open-set (i.e. unknown) speakers, with 5 support samples (only for closed-set) and 5 query samples per speaker. The predictions generated by the model are evaluated using two metrics: closed-set accuracy and open-set AUROC. Closed-set accuracy measures the percentage of correct labels vs. the ground truth on only closed-set query samples, while open-set AUROC measures the model’s ability to differentiate between closed-set and open-set query samples. For the latter metric, we adapted a technique to convert a closed-set classifier for open-set evaluation by taking the ratio between the top two class prediction distances [19, 13], then using the ratio scores to compute the AUROC.

The *speaker pair evaluation* task assesses the model’s performance in accurately classifying positive and negative examples of a single registered speaker. It consists of a list of speaker pairs, and for each speaker pair we sample 10 sequential segments from a random file selected from each speaker. We then compute the score for the speaker pair by averaging the distances between all possible combination of segments ($10 \times 10 = 100$) and use these scores to calculate the equal error rate (EER), defined as where the false acceptance rate and false rejection rate are equal on the ROC curve.

We applied these evaluation tasks individually on VoxCeleb1, BWAD Single, BWAD OE, BWAD IE, and EarSAVAS. For VoxCeleb1, speakers for the few-shot evaluation task were sampled from the entire dataset (1,251 speakers), while for the speaker pair evaluation task we used the VoxCeleb1-E speaker trial list (total of 579,818 speaker pairs). For BWAD and EarSAVAS configurations, speakers for both evaluation tasks

were sampled from the randomly generated test split, and the speaker pair evaluation task was run on 10,000 randomly generated speaker pairs from the test split.

4. Results

Table 1 shows a comparison of our model pre-training and fine-tuning configurations for user authentication on the few-shot and speaker pair evaluation tasks across different test datasets.

4.1. Training Model on Traditional Speech Data Reduces Performance on In-Ear and In-Earcup Mic Data

The pre-trained model without any fine-tuning achieves the best performance on both the few-shot and speaker pair evaluation tasks for VoxCeleb1 compared to other fine-tuned models, and it is comparable to other state-of-the-art models in the current research literature [5]. However, when evaluated on the BWAD and EarSAVAS test splits, the performance of the pre-trained model drops by 4 – 6x for the speaker pair EER and by 26 – 43% for the few-shot open-set AUROC. While the few-shot closed-set accuracy only experiences a minimal drop of 2 – 7%, it does not quantify the model’s ability to reject unknown users; therefore, we believe that the few-shot open-set AUROC and speaker pair EER are more significant metrics to understand the performance implications for our use case in user authentication. In addition, the performance of the pre-trained model is conditional on having good SNR conditions, which will drop significantly in noisy environments or with loud background speech nearby. On the other hand, using inside mics will not suffer from the same problem, as active noise cancellation can typically reduce noise by 20 – 30 decibels.

4.2. Fine-Tuning Model on In-Ear and In-Earcup Mic Data Improves Performance during Evaluation

The model fine-tuned on inside mic data from BWAD Single performed better compared to the pre-trained baseline on all BWAD test splits, with a 22 – 24% improvement in the EER and a 19 – 54% improvement in the few-shot open-set AUROC. Similarly, the model fine-tuned on inside mic data from EarSAVAS performed better than the pre-trained baseline on the EarSAVAS test split, with a significant improvement of 76 – 89% in the few-shot open-set AUROC. However, the improvement in the EER was much smaller (only 3%); this may be due to the frequency response for EarSAVAS looking similar to the outside mic data from VoxCeleb2 (Figure 1d vs. Figure 1a), leading to minimal performance gains. While the fine-tuned models experience a drop in performance on VoxCeleb1 compared to the pre-trained baseline, we consider these results less significant given the varying characteristics of outside vs. inside mic speech data and the goal being to deploy the models on inside mic data.

4.3. In-Ear and In-Earcup Mic Data from Multiple Form Factors Improves Performance

Models fine-tuned on multiple form factors from BWAD performed better than BWAD Single on all BWAD test splits by up to 34% in the EER and by up to 19% in the few-shot open-set AUROC. BWAD All (containing all 4 headphones and earbuds) outperforms the pre-trained baseline by 47 – 50% on the EER and by 41 – 81% on the few-shot open-set AUROC, while BWAD OE performs better than BWAD IE, even when evaluated on the BWAD IE test split. However, BWAD All does not

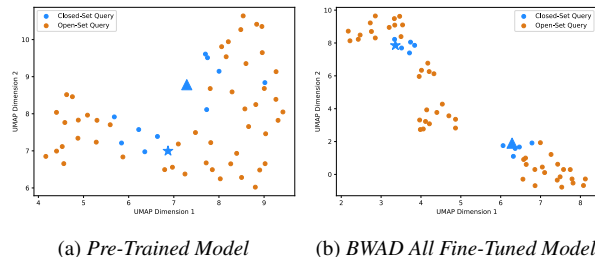


Figure 2: UMAP visualization of the feature embeddings from a 2-way few-shot evaluation task on inside mic data from a single in-ear wireless earbud form factor in BWAD. Dots represent feature embeddings of query set samples, with blue dots from closed-set (i.e. registered) speakers and orange dots from open-set (i.e. unknown) speakers. Triangles and stars represent class prototypes generated from support set samples for each speaker.

outperform EarSAVAS on the EarSAVAS test split, suggesting that in order for the fine-tuning to be effective, data from the target form factor is required.

To qualitatively understand the effect of fine-tuning on inside mic data, we applied UMAP [20] to visualize the feature embeddings generated by our models on BWAD for the 2-way few-shot evaluation task, shown in Figure 2. The pre-trained model embeddings are scattered across the latent space and do not coalesce neatly into clusters, particularly with respect to the closed-set query samples around the speaker prototypes, which will lead to lower performance when differentiating between registered and unknown speakers. In contrast, the BWAD fine-tuned model learns to create tight clusters of the closed-set query samples around the speaker prototypes and better separates the open-set query samples away from the closed-set samples, leading to better performance.

Although the fine-tuned models outperform the pre-trained models on the inside mic data in high SNR conditions, they do not reach the state-of-the-art numbers currently possible for open-set AUROC or EER [21, 22], primarily due to the limited number of speakers in the inside mic datasets (~200 vs. 6,000) and using closed-set models adapted for open-set recognition. This can be improved in the future by training a proper few-shot open-set recognition model and fine-tuning on additional inside mic data, both through collecting data from physical headphones and earbuds or synthesizing it through generative model and signal processing approaches.

5. Conclusion

This paper presents an SID model fine-tuned on speech data recorded from in-ear and in-earcup microphones for user authentication on low-resource headphones and earbuds. In contrast to traditional “outside” microphones, inside microphones capture less background noise and are hard to circumvent through spoofing or other adversarial attack techniques. We evaluate the SID models both with and without fine-tuning and show that fine-tuning leads to improved performance on inside mic data by up to 50 – 81%. In addition, fine-tuning on inside mic data acquired from multiple headphones and earbuds improves the performance of the model, contingent on the form factor of interest being included in the training process.

6. Acknowledgments

We would like to thank all of the individuals who assisted us in the collection of data for the Bose Wearable Audio Dataset.

7. References

- [1] X. Fan, L. Shangguan, S. Rupavatharam, Y. Zhang, J. Xiong, Y. Ma, and R. Howard, "HeadFi: bringing intelligence to all headphones," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 147–159.
- [2] T. Chen, Y. Yang, C. Qiu, X. Fan, X. Guo, and L. Shangguan, "Enabling Hands-Free Voice Assistant Activation on Earphones," in *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*, 2024, pp. 155–168.
- [3] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End Loss for Speaker Verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883.
- [4] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.
- [5] Q. Li, L. Yang, X. Wang, X. Qin, J. Wang, and M. Li, "Towards Lightweight Applications: Asymmetric Enroll-Verify Structure for Speaker Verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7067–7071.
- [6] C. Gao, B. Desplanques, C. J.-T. Ju, A. Chadha, and A. Stolcke, "Post-Training Embedding Alignment for Decoupling Enrollment and Runtime Speaker Recognition Models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 836–10 840.
- [7] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.
- [8] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection," in *Proc. Interspeech*, 2019, pp. 1008–1012.
- [9] F. Rumsey, "Headphone Technology: Hear-Through, Bone Conduction, and Noise Canceling," *Journal of the Audio Engineering Society*, vol. 67, no. 11, pp. 914–919, 2019.
- [10] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4080–4090.
- [11] 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 *Proc. Interspeech*, 2020, pp. 2977–2981.
- [12] J. Wang, K.-C. Wang, M. T. Law, F. Rudzicz, and M. Brudno, "Centroid-based Deep Metric Learning for Speaker Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3652–3656.
- [13] D. Jain, K. Huynh Anh Nguyen, S. M. Goodman, R. Grossman-Kahn, H. Ngo, A. Kusupati, R. Du, A. Olwal, L. Findlater, and J. E. Froehlich, "ProtoSound: A Personalized and Scalable Sound Recognition System for Deaf and Hard-of-Hearing Users," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–16.
- [14] N. Shashaank, B. Banar, M. R. Izadi, J. Kemmerer, S. Zhang, and C.-C. Huang, "HiSSNet: Sound Event Detection and Speaker Identification via Hierarchical Prototypical Networks for Low-Resource Headphones," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [15] X. Zhang, Y. Wang, Y. Han, C. Liang, I. Chatterjee, J. Tang, X. Yi, S. Patel, and Y. Shi, "The EarSAVAS Dataset: Enabling Subject-Aware Vocal Activity Sensing on Earables," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 2, pp. 83:1–83:26, 2024.
- [16] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech*, 2017, pp. 2616–2620.
- [17] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.
- [18] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations (ICLR)*, 2019.
- [19] P. R. Mendes Júnior, R. M. de Souza, R. d. O. Werneck, B. V. Stein, D. V. Pazinato, W. R. de Almeida, O. A. B. Penatti, R. d. S. Torres, and A. Rocha, "Nearest neighbors distance ratio open-set classifier," *Machine Learning*, vol. 106, no. 3, pp. 359–386, 2017.
- [20] J. Healy and L. McInnes, "Uniform manifold approximation and projection," *Nature Reviews Methods Primers*, vol. 4, no. 1, pp. 1–15, 2024.
- [21] S. Huang, J. Ma, G. Han, and S.-F. Chang, "Task-Adaptive Negative Envision for Few-Shot Open-Set Recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 7161–7170.
- [22] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7062–7066.