Speech Recognition as a Threat Vector in Real-Time Doxxing

Beels, C. D.

December 10, 2025

Abstract

This research evaluates the feasibility of speech recognition as a potential threat vector in real-time doxxing operations, using consumer wearable technology devices such as the Ray-Ban Meta Glasses. The study was broken up into two phases: Phase 1, which validated a SpeechBrain [9]-based Python model using audio data from the VoxCeleb2 [7] database, and Phase 2, which used live audio capture from Meta Glasses and compared with a locally-created dataset to examine confidence in identifying a particular individual through speech. Phase 1 produced a 93% hit rate and 100% correct-rejection rate at a cosine similarity threshold of 0.4, with a median processing latency of 0.67 seconds. In Phase 2, the system averaged a 96% correct selection rate between targets in the dataset, however, overall confidence in correct identification remained low (47% and 30%, respectively). These results indicate that speech-based real-time doxxing may be a feasible threat in certain scenarios, particularly with minimal background noise and close proximity between threat actor and target.

1 Introduction

In war, as long as technology continues to develop, adversary strategies will continue to evolve along with it. Nowhere is this more true than in the cyber domain, where any publicly available technology could pose a future threat. This research seeks to understand the modern evolution of a longstanding threat within cyberspace: **doxxing**. Douglas informs us that the term "doxxing" originates from 1990's hacker culture, as a shortened version of "dropping dox." [1] It signifies the public release of a person's documents or identifying data. He further distinguishes between three forms of doxxing: 1. **Deanonymizing:** When a previously anonymous individual is identified, 2. **Delegitimizing:** When information is released to deliberately shame or debase an individual, and 3. **Targeting:** When information is released to locate someone or their home. [1]

To this day, doxxing poses a threat to operation security (OPSEC) for U.S. Military personnel; however, it is traditionally a slow and deliberate process that

cannot happen in real-time using live data. This is changing with the advent of **real-time doxxing**, which utilizes the live capture and analysis of data on an individual to expose their identity in the moment.

Recent advances in consumer wearables, particularly the Meta Ray-Ban glasses capable of hands-free audio/video capture, create a plausible threat vector for doxxing in real-time at close range. Two Harvard students showcased that the possibility of a threat actor conducting real-time doxxing was legitimate in their 2024 project *I-XRAY* [4]. Their project shows that ordinary people—and therefore enemy forces—can pose a legitimate threat to operation security using consumer wearables and modern methods of real-time doxxing.

This paper focuses on the feasibility of live speech recognition as a method of real-time doxxing. This research is conducted in conjunction with an investigation into facial recognition, with the purpose of audio collection being an additional By showing the practical efficacy of audio data collection using the Meta Ray-Ban glasses, we can more clearly determine what specific threat vectors must be considered, thereby informing defensive actions in a threat-likely scenario.

1.1 Gap in Research

Previous projects, namely *I-XRAY*, have demonstrated the utility of imaging and facial recognition in real-time doxxing [4], however this will not always be reliable due to various environmental factors. Voice is a promising alternative metric that few researchers have deeply considered. There is no current wealth of knowledge on real-time audio-based doxxing. This project will consider the efficacy of speech analysis as a secondary method of real-time doxxing by building a live prototype using Meta Ray-Ban glasses in conjunction with the SpeechBrain Python library and testing in simulated conditions. The results of this project seek to demonstrate speech as a potential threat to OPSEC.

1.2 Research Question

Central to this research is answering the following question: Can processing live audio captured using Meta Ray-Ban glasses meaningfully increase the likelihood of positive target identification in real-time doxxing while maintaining latency and accuracy constraints relevant to a theoretical real-life scenario?

This study serves as an experimental proof-of-concept. It will support future examination of various threat vectors, and development of response protocol protecting against real-time doxxing.

1.3 Related Works

Snyder et al. (2017) provides a novel approach to quantitatively describing doxxing. Their research returns some significant results, namely that the most cited justifications for doxxing are “justice” and “revenge,” and that efforts by

large organizations to mitigate doxxing and prevent its use have been largely successful in reducing its prevalence. [3]

When discussing the usage of speech recognition in doxxing, it is first necessary to bring up the foundations of modern audio recognition and recognize the tools which are being utilized in this project. Hinton et al. (2012) demonstrated that deep neural networks were useful in automatic speech recognition (ASR) applications and were more effective than the previously-used Gaussian mixture models. [2] Since then, deep learning has been the standard for ASR, and projects such as PyTorch, introduced by Paszke et al. (2019) and SpeechBrain, introduced by Ravanelli et al. (2021) built off this knowledge to make deep-learning based ASR into an open-source framework, which is the backbone for the real-time doxxing speech pipeline that was analyzed in this study. [8], [9]

Contemporary projects are already showing that real-time doxxing can be effective in certain environments. One highly popular example is the 2024 Harvard student project, *I-XRAY*, which demonstrated the potential for real-time doxxing by utilizing Meta Ray-Ban glasses in conjunction with the facial search service PimEyes. [4] Although the project was intended as an academic demonstration, it shows that modern consumer devices like the Meta glasses can be used to expose peoples' identities at a relatively low cost and effort.

2 Methodology

This study was conducted in two phases. Phase 1 was a verification of the speech recognition program using an open source speech dataset known as *VoxCeleb2*. VoxCeleb2 was selected for its vast number of subjects with high-quality data, and its anonymity, which is necessary to avoid unwanted harm from doxxing. This was done to ensure our code was effective for a variety of voice profiles and acoustic conditions, and to get a baseline level of system performance.

Phase 2 was a live implementation of the code using the Meta Ray-Ban glasses, intended to simulate a real-time doxxing attempt. This phase was conducted to better understand performance of our adversarial system, and the constraints that hardware brings in latency and accuracy. This structure of methods enabled greater replicability by not relying solely on collected data for verification of the code.

2.1 Model Overview

This project utilized SpeechBrain [9], a Python and PyTorch [8]-based speech recognition toolkit, to determine the identity of a given speaker from a given dataset. The model captured data using either a database (VoxCeleb2) or using live audio collection (Meta Ray-Ban glasses), and processed the data through a program in Jupyter Notebook. The MatPlotLib library was used for data visualization.

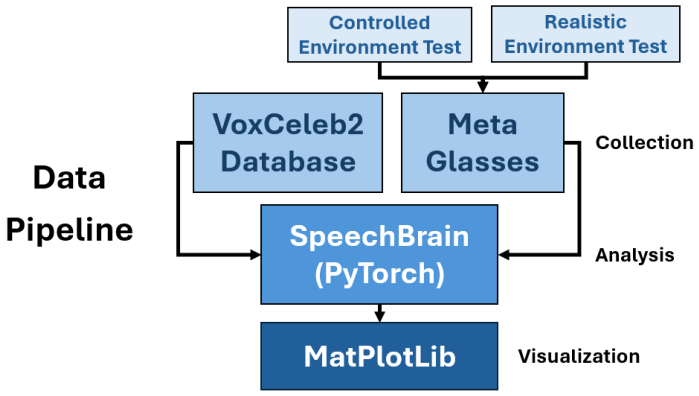

Figure 1: Data Pipeline Flowchart

2.2 Model Preparation

For Phase 1, audio was sourced from the VoxCeleb2 database, which includes real-world voice recordings of public figures in a variety of environments. [7] 4 speakers were chosen from the VoxCeleb2 dataset as doxxing targets. For each target, 10 audio clips were chosen at random as training data, and another 25 random clips were chosen as testing data. An additional 100 audio clips were selected at random from the dataset as false positives, to determine if the model could both correctly identify a speaker and correctly reject a non-target. Audio clips were segmented into 5 to 10 second chunks in order to better represent voice capture for real-time doxxing.

For Phase 2, audio from 3 different individuals was recorded locally as reference data in roughly 5 second clips. One individual was designated as the target, and the other two as false-positives.

Audio clips were pre-processed prior to use with SpeechBrain. Clips were altered to use a sampling rate of 16kHz mono in order to work with the ECAPA-TDNN [6] speaker-embedding model used by SpeechBrain, and generate an embedding for each audio clip. After pre-processing, clips were stored in a new directory labeled with the corresponding individual's id number.

2.3 Model Evaluation

For both phases, 10 clean audio clips were used to generate reference embeddings. Then the testing data (200 VoxCeleb2 audio clips for Phase 1 and Meta Ray-Ban audio capture for Phase 2) were passed through the ECAPA-TDNN model to get an experimental embedding. The cosine similarity between the reference and experimental embeddings were computed, giving a verification score.

In order to determine the most effective verification score, the model was

tested with various cosine similarity thresholds, ranging from 0.3 to 0.8, and the threshold with the greatest success rate was chosen.

If the experimental score was greater than the verification score, then the program labeled that audio as a "match," and for all other cases output "no match."

For the live test in Phase 2, audio was collected continuously via the onboard microphone embedded into the legs of the Meta glasses. The Meta Ray-Ban glasses were connected via Bluetooth to a mobile phone. In order to run the Python code live, a base computer hosted Jupyter Notebook at a HTTP port, which the mobile phone connected to. Audio data was processed remotely on the computer.

An example of a result from the test can be seen in Figure 2, showing the output in Jupyter Notebook. Since the cosine similarity, labeled as "sim," is below the example verification score of 0.5, the voice is listed as unknown.

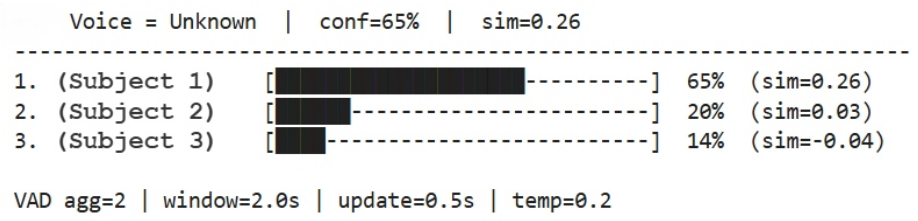

Figure 2: Speech Identification Output Example

2.3.1 Metrics

The performance of this model was evaluated on the following metrics:

1. Accuracy: Proportion of program run time where the speaker was correctly identified.
2. Latency: Processing time for each recording chunk.

3 Results

3.1 Phase 1

First, it was necessary to find an appropriate cosine similarity threshold for distinguishing between the four categories of audio classification:

- Hit: The audio was from a target, and the model identified the correct target.
- Miss: The audio was from a target, but the model identified either the wrong target or no target.

- False Positive: The audio was not from a target, but the model identified a target.
- Correct Rejection: The audio was not from a target, and the model did not identify a target.

As stated before, the model for Phase 1 was tested with cosine similarity thresholds ranging from 0.3 to 0.8. This test is summarized in the table depicted by Figure 3:

	threshold	hits	misses	false_positives	correct_rejections	match_acc	nomatch_acc	overall_acc	false_positive_rate
0	0.3	73	27	1	99	0.73	0.99	0.860	0.01
1	0.4	93	7	0	100	0.93	1.00	0.965	0.00
2	0.5	83	17	0	100	0.83	1.00	0.915	0.00
3	0.6	43	57	0	100	0.43	1.00	0.715	0.00
4	0.7	3	97	0	100	0.03	1.00	0.515	0.00
5	0.8	0	100	0	100	0.00	1.00	0.500	0.00

Figure 3: Cosine Similarity Thresholds

The threshold of 0.4 was chosen due to its highest overall accuracy rate, denoted as "overall_acc," and was used in all further tests. Next, the 4 classifications were plotted by re-running the model with 200 trials at a threshold of 0.4, shown in Figure 4.

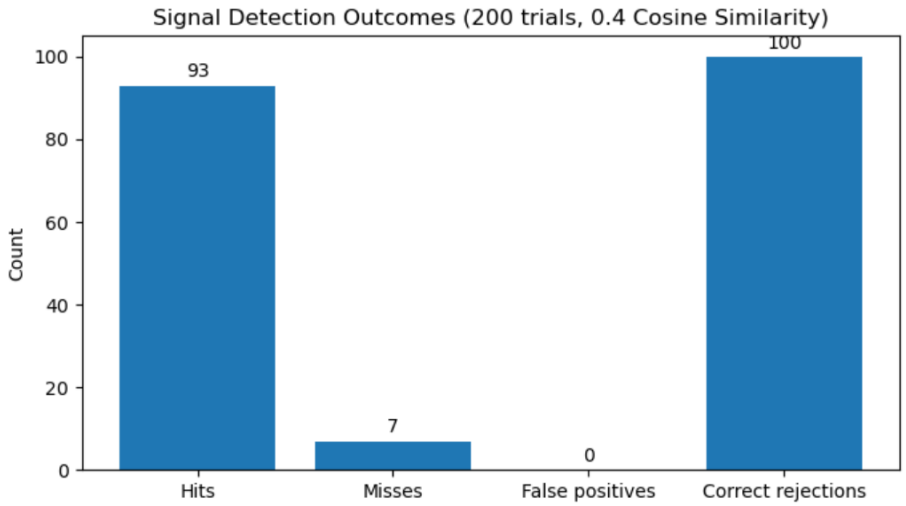

Figure 4: Signal Detection Outcomes

The model had a 93% hit rate, and a 100% correct rejection rate, resulting in a 96.5% success rate overall. Among the 100 positive audio clips, the hit rate is broken down between the 4 targets as is shown in Figure 5.

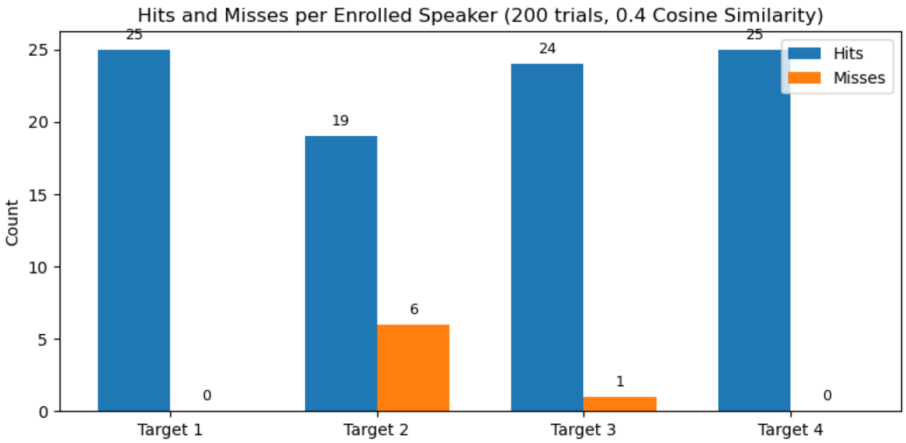

Figure 5: Hit Rate per Target Speaker

Lastly for Phase 1, the latency, or time taken to analyze each audio clip, was plotted as a histogram in Figure 6. The mean latency for one clip was 0.84 seconds, median latency was 0.67 seconds, and 95% of audio clips were processed in less than 1.8 seconds.

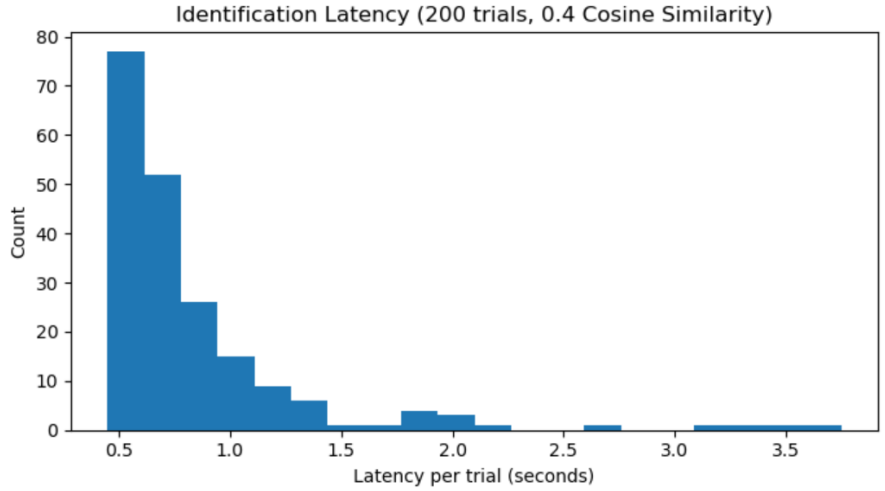

Figure 6: Latency per Audio Clip

3.2 Phase 2

In Phase 2, two tests were ran. Test 1 was done using the Meta Ray-Ban glasses in a controlled environment with no deliberate background noise. Test 2 was done in an environment with deliberate background noise played over the Target speaking. The way this noise was generated was by using a speaker, playing "Busy Airport Sounds" at a moderate volume placed 1 meter away from the wearer of the Meta glasses. In both cases, the target was speaking at a typical conversation volume standing 1 meter away from the Meta glasses as well. Both tests were conducted for roughly 200 seconds, sampling every 2 seconds.

In both tests, the model was confident that the speaker was most likely Target 1 of the 3 targets, which was the correct target, but relatively uncertain that the speaker was in fact Target 1. As shown in Figure 7, With no background noise, our best match accuracy was 94%, but the final verdict had only a confidence rate of 47%. With background nose, the best-match accuracy was 98%, but only 30% of samples were certain in their target identification.

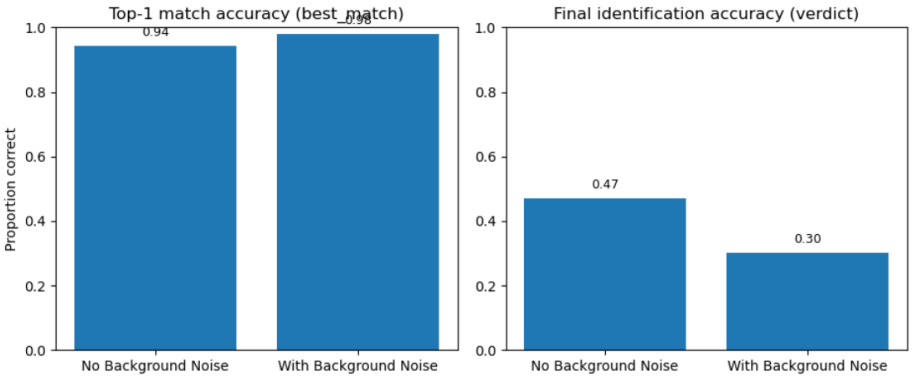

Figure 7: Best-Match and Final Verdict Accuracy

The distribution of measured cosine similarities of each sample were graphed in Figure 8, distinguishing between the control test with no background noise and the test with background noise. Between the two tests, 50 of the 106 samples from Test 1 and 30 of the 99 samples from Test 2 had cosine similarities greater than 0.4.

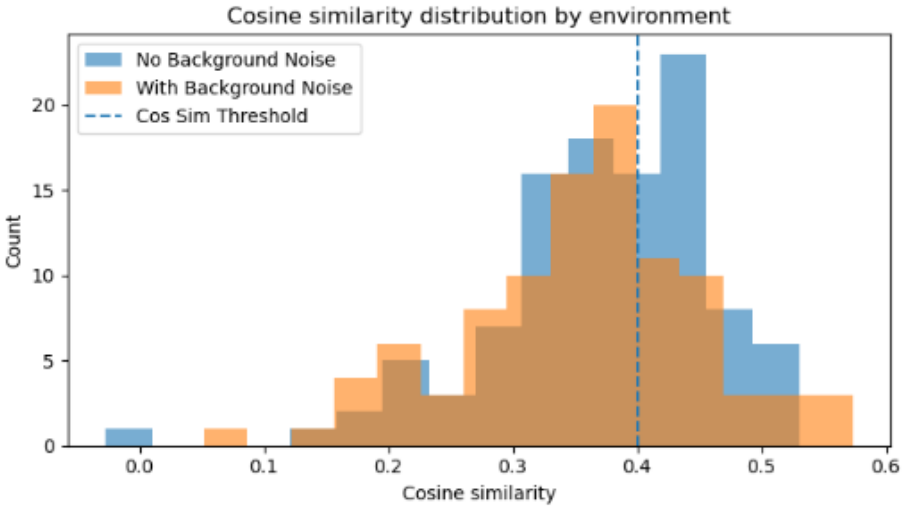

Figure 8: Cosine Similarity Distributions

Finally, the latency rate for Phase 2 was a constant 0.5 seconds for the model to update to its current verdict in both tests.

4 Discussion

As aforementioned, the question this research sought to evaluate has been: Can processing live audio captured using Meta Ray-Ban glasses meaningfully increase the likelihood of positive target identification in real-time doxxing while maintaining latency and accuracy constraints relevant to a theoretical real-life scenario?

From Phase 1, our model validation went very well, with a high hit rate of 93% and a 100% correct rejection rate (Figure 4). Our model latency also had a promising distribution centering around 0.7 seconds, with some variation resulting from variations in audio clip file size and length (Figure 5). With a controlled audio input such as VoxCeleb2, ASR has a high likelihood of success.

When transitioning to Phase 2, model confidence decreased significantly. The model did have a very high accuracy in distinguishing among the three targets in the dataset, however its confidence in having identified Target 1 was low, around 47% for the control group (no background noise) and 30% with noise (Figure 6). This shows the model struggled with getting a high cosine similarity, and thus a high confidence, in a more realistic scenario. The current level of confidence would likely not have given an adversary a useful level of certainty that they have positively identified their target. In a scenario where background noise would be even higher, one would expect these results to decline and become even less effective.

One area that did see significant success was in managing latency. A concern before conducting live testing was that streaming audio to a HTTP port would cause a delay, or that latency would cause issues in some regard, but what was observed was that with good connectivity, latency rarely caused greater than 1-2 seconds of delay between audio capture and verdict declaration.

4.1 Limitations

One limitation of this study is that the audio from VoxCeleb2 used in phase 1 does not map 1-1 to the live audio capture in phase 2. Microphone quality, compression standards, and other variables likely caused discrepancies in the ways in which audio was processed by SpeechBrain, leading to some of the difference in confidence seen between the two phases.

The live sample size of only 3 people is another clear limitation. One may argue it reflects the operational restrictions for the enemy in a realistic scenario, however it inherently also limits the generalization of the study.

This study assumes that the artificially-induced background noise is representative of background noise in general, when this in truth may not be the case. Future work that expands on this threat vector will certainly benefit from having more test subjects and more data, but may also benefit from trying different kinds of noise generation to model various different scenarios.

5 Conclusion

This study provides evidence that the use of speech recognition with consumer wearable devices as a method of real-time doxxing is at least partially feasible. As live tests proved to have a moderate to high level of accuracy in a controlled environment, USF may wish to consider withholding speech in relatively quiet situations, such as the waiting areas of an airport. This study does also support that latency and audio processing time would likely not be a barrier stopping real-time speech recognition. This research especially supports potential threat calculations for special and covert operations, which may need to consider not only when, but *how* the enemy is listening.

While previous research has built the speech recognition tools and frameworks this study utilizes (SpeechBrain, PyTorch, the VoxCeleb2 dataset), and projects like *I-XRAY* have proven that facial recognition is a threat vector, this study examines a novel combination of these concepts and shows that with off-the-shelf items, a real-time audio-based doxxing pipeline is a feasible threat.

As with the rest of the cyber domain, wearable consumer technology will only continue to develop in the future, and new threat vectors will continue to emerge in cyberspace. It is up to the United States to remain adaptable and ready for the next new threat, no matter the form it may come in.

References

[1] D. M. Douglas, "Doxxing: A conceptual analysis," *Ethics and Information Technology*, vol. 18, no. 3, pp. 199–210, 2016. [Online]. Available: `https://link.springer.com/article/10.1007/s10676-016-9406-0`

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[3] P. Snyder, P. Doerfler, C. Kanich, and D. McCoy, "Fifteen minutes of unwanted fame: Detecting and characterizing doxxing," in *Proc. 2017 Internet Measurement Conf. (IMC '17)*, Nov. 2017, pp. 432–444. doi:10.1145/3131365.3131385

[4] Harvard Law School Library Innovation Lab, "I-XRAY: the AI glasses that reveal anyone's personal details" 2024. [Online]. Available: `https://lil.law.harvard.edu/events/i-xray-lunch/`

[5] M. Ravanelli *et al.*, "Open-Source Conversational AI with SpeechBrain 1.0," *arXiv*, 2024. doi:10.48550/arXiv.2407.00463

[6] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN-based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020. [Online]. Available: `https://doi.org/10.48550/arXiv.2005.07143`

[7] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.

[8] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019, pp. 8024–8035.

[9] M. Ravanelli *et al.*, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.