



# Adversarial Attacks on Text-dependent Speaker Verification System

Sreekanth Sankala, Venkatesh Parvathala, Ramesh Gundluru, Sri Rama Murty Kodukula

Department of Electrical Engineering, Indian Institute of Technology Hyderabad, INDIA  
{ee20resch11011, ee22resch01005, ee22m24p000001}@iith.ac.in, ksrm@ee.iith.ac.in

## Abstract

Adversarial attacks against text-independent speaker verification (TI-SV) systems assume access to genuine speaker's enrollment speech ( $e[n]$ ). This assumption is self-defeating because if an attacker has  $e[n]$ , they can bypass the system directly, making adversarial examples unnecessary. In contrast, we observe that the text-dependent SV (TD-SV) system, where the genuine speaker must say a password, offers a more practically relevant attack scenario. In reality, the attacker may not have access to the password spoken by a genuine speaker, but they can likely obtain normal speech from the genuine speaker. Therefore, generating adversarial noise that, when added to the genuine speaker's normal speech, can bypass the password requirement of a TD-SV system constitutes a potential realistic attack. This work investigates the feasibility of such a practical attack and shows that even the state-of-the-art TD-SV system is vulnerable with an attack success rate of 64.28 %.

**Index Terms:** Speaker verification, Adversarial attacks, Text-dependent speaker verification, Real-world adversarial attacks.

## 1. Introduction

Speaker verification (SV) is the task of authenticating the claimed identity of a speaker from his/her voice characteristics. Based on the mode of verification, SV systems are divided into two types: 1) TI-SV and 2) TD-SV. The TI-SV system verifies only the speaker's characteristics, while the TD-SV system verifies both the speaker's characteristics and the textual content of the speech signal (password). With the advent of deep learning techniques, both the TI-SV and TD-SV systems have shown significant performance improvements on various benchmark datasets [1]. Given their real-world deployment [2, 3], it is important to investigate the potential attacks against them.

There are two primary types of attacks against SV systems: 1) speaker-specific attacks and 2) model-specific attacks. Speaker-specific attacks attempt to replicate the voice characteristics of the genuine speaker, often using voice conversion (VC), text-to-speech (TTS) systems, etc. These attacks have been extensively studied under the category of spoof attacks through various automatic speaker verification spoof challenges [4–6]. In contrast, model-specific attacks, which seek to exploit model vulnerabilities to alter SV system decisions, are relatively less explored [7–9]. This work focuses on adversarial attacks, a type of model-specific attack, in the context of SV systems.

A typical text-independent speaker verification (TI-SV) system takes the genuine speaker's enrollment speech ( $e[n]$ ) and a test speech signal ( $t[n]$ ) as input and determines if the test speech is from the genuine speaker. The standard approach computes a speaker similarity score ( $f_{\theta}\{e[n], t[n]\}$ ) between  $e[n]$  and  $t[n]$ , and compares it to a threshold for verification

[10]. Ideally, a TI-SV system outputs a higher score if  $t[n]$  is from the genuine speaker, and a lower score if  $t[n]$  is from an imposter. An adversarial attack against the TI-SV system involves slightly perturbing the  $t[n]$  to overturn the verification result, i.e., accept an imposter speaker or reject the genuine speaker [8, 11], as shown in Figure 2. Studies on adversarial attacks against TI-SV systems showed that even state-of-the-art TI-SV systems are vulnerable to adversarial attacks [7–9, 12]. These studies assume that the adversary knows the TI-SV system  $f_{\theta}(\cdot, \cdot)$  (parameterized by  $\theta$ ) and has access to both the genuine speaker's enrollment speech ( $e[n]$ ) and the imposter's test speech ( $s[n]$ ). It then computes adversarial noise  $v[n]$  using partial derivatives of  $f_{\theta}(e[n], s[n])$  with respect to  $s[n]$  (as detailed in Section 2) [8]. Adding this adversarial noise ( $v[n]$ ) to the imposter's speech  $s[n]$  causes the TI-SV system to incorrectly accept it as genuine. These types of attacks are referred to as gradient based white-box adversarial attacks in the literature.

While the gradient-based white-box adversarial attacks are powerful, they encounter numerous challenges when launched in real-world scenarios [13]. A key challenge is accessing the target system, in this work it is the TI-SV system  $f_{\theta}(\cdot, \cdot)$ . A common approach is to build a surrogate system to approximate the target system [14], and use it to generate adversarial noise. These attacks, known as transferable black-box attacks, are effective when the surrogate system closely approximates the target [15]. However, even the transferable black-box attacks against the TI-SV system assume the attacker has access to both the genuine speaker's enrollment speech ( $e[n]$ ) and the imposter speaker's test speech ( $s[n]$ ). Most studies on adversarial attacks against TI-SV systems rely on this assumption to explore the TI-SV system's vulnerabilities [7–9, 13, 16–18].

In this work, we revisit these assumptions and identify an unrealistic one. In the case of the TI-SV system, if the adversary has access to  $e[n]$ , it can be used directly for access, rendering the generation and addition of adversarial noise  $v[n]$  to the imposter speaker's voice  $s[n]$  unnecessary. This suggests that the current setup for adversarial attacks against TI-SV systems is unrealistic. Despite their limitations, studying these attacks can offer insights into TI-SV system functionality and its vulnerability to such small perturbations, ultimately aiding the development of more robust systems [19].

In contrast to the attacks against TI-SV systems, we observe that adversarial attacks against TD-SV systems are relatively more realistic. TD-SV systems require the correct password spoken by the genuine speaker. Attackers typically cannot access the genuine speaker's spoken password. However, natural conversations with that speaker are often available. If an attacker can use these conversations to generate adversarial noise and add it to the genuine speaker's normal speech to bypass the password verification by the TD-SV system, this constitutes a

realistic attack. The main objective of this paper is to study the feasibility of such a realistic attack against TD-SV systems.

The remainder of the paper is organized as follows: Section 2 outlines the proposed method for attacking TD-SV systems. Section 3 details the dataset and experimental setup. Section 4 examines the feasibility of attacks against TD-SV systems. Finally, Section 5 outlines the planned future work.

## 2. Adversarial Attacks Against SV System

### 2.1. TD-SV system

TD-SV systems compare two variable-length speech signals to determine if they belong to the same speaker and contain the same password information [20]. Given the challenges associated with directly comparing variable-length speech signals, the predominant approach involves deriving fixed-dimensional embeddings that are specific to the speaker and the password present in the signal [21, 22]. The extracted fixed-dimensional embeddings are then compared using a simple similarity measure, such as cosine similarity, to obtain a similarity score between the two signals. Finally, the similarity score is compared against a threshold for the binary decision, i.e., accept/reject.

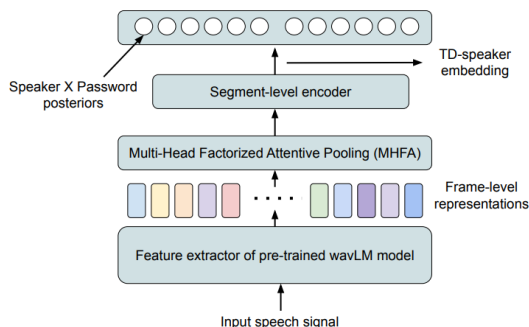


Figure 1: Block diagram of text-dependent speaker classifier.

#### 2.1.1. Extraction of fixed dimensional embeddings

Deep neural networks are now widely used for extracting fixed-dimensional embeddings [21, 22]. These methods train a classification network on a large dataset of background speakers and passwords to simultaneously identify both. Once the classification network is trained, the penultimate layer’s activations are extracted to obtain a fixed-dimensional embedding representative of speech signal, as shown in Figure 1.

The performance of deep embedding-based TD-SV systems depends on the size of the training data. However, TD-SV training datasets are often limited because data containing real passwords are not shared due to privacy concerns [23, 24]. To tackle the data scarcity issue, pretrained self-supervised models are employed as front-end feature extractors and are jointly fine-tuned with the classification network. [25–29]. Figure 1 shows the text-dependent classifier used in this work. In this setup, the pretrained WavLM [30] model acts as the front-end feature extractor. Variable-length representations extracted by this front-end feature extractor are pooled using a Multi-Head Factorized Attentive (MHFA) pooling layer to produce a fixed-dimensional embedding. This fixed-dimensional embedding is then processed through a couple of dense layers in the segment-level encoder to produce speaker-password posteriors. The weights of the network are optimized to min-

imize the additive angular margin softmax loss [31]. More details about the network architecture can be found in [32]. While the work in [32] uses it as a text-independent classifier that only classifies speaker labels, this work employs it as a text-dependent speaker classifier that distinguishes both speaker and password labels. In other words, for a TI classifier, the number of classes equals the number of speakers, while for a TD classifier, the number of classes is the product of the number of speakers and the number of passwords. The pre-trained WavLM model used in this work is publicly available at <https://github.com/microsoft/unilm/tree/master/wavlm> under the name “WavLM Base+.”

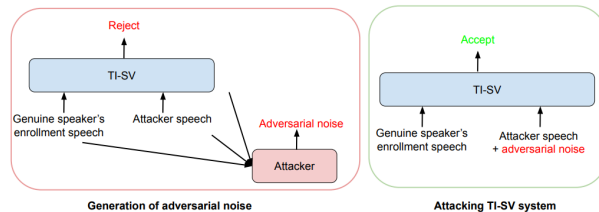


Figure 2: Adversarial attacks against TI-SV system

### 2.2. Adversarial attacks against TD-SV system

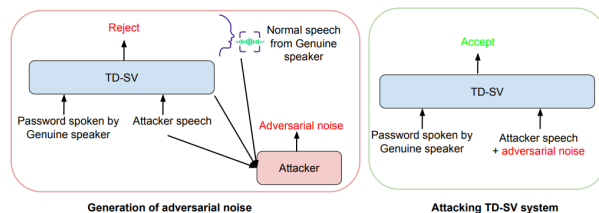


Figure 3: Proposed setup of adversarial attack against TD-SV system. Key novelty is that the attacker does not require the genuine speaker’s speech containing the correct password.

The TD-SV system compares  $e[n]$  and  $t[n]$  to assess whether they come from the same speaker and contain the same password. A typical TD-SV system commonly operates on four different types of test speech signals, as shown in Table 1. An ideal TD-SV system accept only TC and reject the other three types (TW, IC, IW) of test speech signals. As shown in Figure 3, an adversarial attack against the TD-SV system involves adding adversarial noise  $v[n]$  to the attacker sentence such that the system incorrectly accepts the modified attacker sentence. In this work, we propose a new method for generating adversarial noise ( $v[n]$ ) for TD-SV systems. The key distinction is that, unlike the attack against TI-SV systems, our approach does not require genuine speaker’s enrollment speech (password spoken by genuine speaker in TD-SV system) to create the adversarial noise. Figure 3 shows the pictorial illustration of the proposed setup of adversarial attack against the TD-SV system.

#### 2.2.1. Generation of adversarial noise ( $v[n]$ )

Consider  $M$  utterances sourced from the natural conversations of the genuine speaker. We refer to these  $M$  utterances as Target Wrong (TW) utterances because they do not include the correct password of a genuine speaker. Let us denote these  $M$  utterances with  $\{s_{tw}^i[n]\}_{i=1}^M$ , where  $s_{tw}^i[n]$  represents the  $i$ -th target

Table 1: Types of test speech signals  $t[n]$  defined by how they are compared to enrollment speech signal  $e[n]$  in TD-SV system.

Type of $t[n]$	Speaker match	Password match
Target Correct (TC)	✓	✓
Target Wrong (TW)	✓	✗
Imposter Correct (IC)	✗	✓
Imposter Wrong (IW)	✗	✗

wrong utterance and  $M$  is the total number of such utterances. Let's consider a attacker utterance  $s[n]$  that is spoken by the attacker. The goal is to add adversarial noise  $v[n]$  to the attacker utterance  $s[n]$ , so that  $s[n] + v[n]$  gets accepted by the TD-SV system. We propose to minimize the following objective function to generate the adversarial noise ( $v[n]$ ).

$$\mathcal{L}(v[n]) = \sum_{i=1}^M \max(\tau - f_{\theta}(s_{tw}^i[n], s[n] + v[n]), -k) \quad (1)$$

where  $f_{\theta}(s_{tw}^i[n], s[n] + \hat{v}[n])$  is the similarity measure between  $i^{th}$  target wrong utterance ( $s_{tw}^i[n]$ ) and adversarial noise added attacker utterance ( $s[n] + \hat{v}[n]$ ).  $f_{\theta}(\cdot, \cdot)$  is TD-SV system parameterized by the model parameters  $\theta$ .  $\tau$  is the threshold used by the TD-SV system to compare the similarity measure and make the final binary decision. " $k$ " is the attack confidence.  $M$  is the total number of target wrong utterances. Note that  $\mathcal{L}(v[n])$  is minimum only when

$$f_{\theta}(s_{tw}^i[n], s[n] + v[n]) > \tau + k \quad \forall i \quad (2)$$

By minimizing the objective function  $\mathcal{L}(v[n])$ , we obtain adversarial noise  $v[n]$ . Adding this adversarial noise to the attacker utterance aims to make it closer to all  $M$  target wrong utterances. We hypothesize that if  $M$  is large enough to encompass a wide range of passwords, this manipulated attacker utterance will also match the target correct utterance. We follow the standard gradient descent algorithm to find out the  $v[n]$  that minimizes the objective function  $\mathcal{L}(v[n])$ . In specific, the adversarial noise  $v[n]$  is initially set with small Gaussian noise and is iteratively updated using the Equations 3 and 4.

$$v[n]_{new} = v[n]_{old} - \alpha \cdot \text{sign}(\Delta_{v[n]} \mathcal{L}(v[n])) \quad (3)$$

$$v[n]_{new} = \text{Clip}_{\epsilon}(v[n]_{new}) \quad (4)$$

Here,  $\alpha$  represents the learning rate.  $\Delta_{v[n]} \mathcal{L}(v[n])$  is the partial derivative of loss function with respect to the  $v[n]$ .  $\text{sign}(\cdot)$  function outputs 1 for positive inputs and -1 for negative inputs. To ensure that the generated adversarial noise remains within a small norm, constraints are imposed on it. Specifically, the  $\text{Clip}_{s[n], \epsilon}(\cdot)$  operation is used to constrain the adversarial noise to be within the range  $[-\epsilon, \epsilon]$ . This is equivalent to enforcing  $l_{\infty}$  norm constraint on the generated adversarial noise, i.e.,  $|v[n]|_{\infty} \leq \epsilon$ . After  $N$  iterations of the update process described in Equations 3 and 4, the algorithm generates  $v^N[n]$ , which we define as the adversarial noise  $v[n]$  in this work.

### 3. Datasets & Experimental details

The TD-SV system in this work is trained using the Track 1 training data from the TDSV 2024 challenge [33]. This Track 1 training data consists of multiple repetitions of 10 passwords,

**Algorithm 1** Evaluating the average attack success rate

---

```

1: for  $spk = 1$  to number of speakers (62) do
2:   for  $pwd = 1$  to number of passwords (10) do
3:     Get  $N$  target correct utterances:  $s_{tc}^i[n]^N$ 
4:     Get  $M$  target wrong utterance:  $s_{tw}^i[n]^M$ 
5:     Get 1 attacker sentence:  $s[n]$ 
6:     Generate adversarial noise:  $v[n]$ 
7:     Generate adversarial signal:  $a[n] = s[n] + v[n]$ 
8:     Attack the system using adversarial signal ( $a[n]$ )
9:     success = 0
10:    for  $i = 1$  to  $N$  do
11:      score =  $f_{\theta}(s_{tc}^i[n], a[n])$ 
12:      val =  $\mathbb{1}(\text{score} \geq \tau)$   $\tau$  is threshold
13:      success = success + val
14:      Attack_success_rate =  $\frac{\text{success}}{N}$ 
15: Average attack success rate =  $\frac{\text{Attack\_success\_rate}}{620}$ 

```

---

5 English and 5 Persian, spoken by 1620 speakers. Therefore, the TD-SV classifier is trained using this speech data, with the number of classes set at 16,200 (1620 speakers x 10 passwords). Training strategies, including loss function, optimizer, learning rate, and learning rate scheduler, follow the approach in [34]. After training the TD-SV classifier, the activation potentials of the first dense layer in the segment-level encoder are recorded to obtain the TD speaker embeddings, as shown in Figure 1. During the verification phase, TD-speaker embeddings extracted from  $e[n]$  and  $t[n]$  are compared using the cosine similarity score to obtain the overall similarity between  $e[n]$  and  $t[n]$ .

The performance and adversarial vulnerability of the TD-SV system are evaluated using the RedDots database, which comprises four evaluation sets: Common Pass-phrases Text-Dependent, Unique Pass-phrases Text-Dependent, Free-choice Pass-phrases Text-Dependent, and Text-Prompted. This work focuses on the Common Pass-phrases Text-Dependent evaluation set, which includes 10 identical passwords spoken by 62 different speakers, consisting of 49 males and 13 females.

#### 3.1. Experimental setup of adversarial attacks against TD-SV system

The RedDots evaluation data includes multiple repetitions of 10 passwords spoken by 62 speakers. This dataset is utilized to examine the adversarial vulnerability of the TD-SV system. To simulate a practical scenario for conducting adversarial attacks against TD-SV systems, we employed the following strategy. For each speaker and password, target wrong utterances and one imposter wrong utterance are assigned. Target wrong utterances are obtained from the target speaker using the remaining nine passwords. The imposter wrong utterance is generated by randomly selecting an imposter speaker and having them use one of the nine remaining passwords. Adversarial noise  $v[n]$  is created using the target wrong utterances and added to the imposter wrong utterances. The resulting signal, which is the imposter wrong utterance with added adversarial noise, is compared to the target correct utterances (multiple repetitions of the target speaker saying the correct password). The effectiveness of the attack is measured by the percentage of these comparisons that are accepted by the TD-SV system. We apply this procedure to all speakers and passwords and compute the average attack success rate, as described in Algorithm 1.

## 4. Results & Analysis

### 4.1. Performance of TD-SV system

Table 2: *Equal Error Rate (EER) of TD-SV system*

System	Type of test speech signal			Average
	TW	IC	IW	
i-vector/GMM [35]	<b>0.43</b>	2.07	3.76	2.08
i-vector/HMM [35]	1.11	1.88	0.46	1.15
wavLM	2.22	<b>0.46</b>	<b>0.18</b>	<b>0.95</b>

Table 2 reports the performance of TD-SV system evaluated on RedDots evaluation data (task m-part-01), as per protocols in [23]. Table 2 shows that wavLM-based TD-SV system achieves lower average EER, computed by averaging the EERs of three different test set (TW, IC, IW) evaluations. Given the performance of wavLM based TD-SV system and their widespread use in downstream tasks (e.g., speaker verification, language identification, emotion recognition) [30, 34, 36], this study investigates the adversarial vulnerability of TD-SV systems using the wavLM-based system described in Section 2.1.

### 4.2. Adversarial vulnerability of TD-SV system

Adversarial noise  $v[n]$  is added to the attacker sentence and compared against  $N$  repetitions of target correct (password spoken by genuine speaker) utterances. Attack success rate is the percentage of these comparisons accepted by the TD-SV system. Algorithm 1 details the calculation of the average attack success rate using data from 62 speakers and 10 passwords, sourced from part 1 of the RedDots database.

Table 3: *Average attack success rate and the average signal-to-noise ratio (SNR) at which the adversarial noise is added.*

Attacker sentence	Attack success rate (%)	SNR (dB)
Target wrong	64.28	30.30
Imposter correct	60.52	35.01
Imposter wrong	42.12	32.37

Table 3 presents the average attack success rate computed with our algorithm with different attacker sentences. The attacker sentence differ from the genuine speaker’s spoken password by speaker information, password information, or both. Therefore, the adversarial noise added to the attacker sentence must incorporate the missing speaker or password characteristics. Specifically, for target wrong sentences, the noise incorporates the missing password characteristics; for imposter correct sentences, the missing speaker characteristics; and for imposter wrong sentences, both speaker and password characteristics.

The results in Table 3 demonstrate that the proposed adversarial attack algorithm is effective at attacking the TD-SV system. In specific, an attack success rate of 64.28 % is reported with target wrong sentence as attacker sentence, even at the addition of adversarial noise at 30.30 dB SNR. It indicates that the addition of adversarial noise to the target-wrong utterance is able to incorporate the password characteristics learned by the TD-SV system. This experimental setup and its resulting attack success rate demonstrate the feasibility of real-world attacks against the TD-SV system. Examples of adversarial signals generated using the proposed method are available

at the GitHub link: <https://speechpublications.github.io/interspeech2025.html>

We also investigated a scenario where the attacker knows the genuine speaker’s password but not their spoken password. The attacker can then generate an imposter correct utterance by speaking the correct password. Our proposed attack algorithm adds adversarial noise to this utterance to incorporate the missing speaker characteristics. This approach achieved a 60.52 % attack success rate. Finally, we considered the case where the attacker lacks both speaker characteristics and password information. Here, the attacker uses an incorrect password to generate imposter wrong utterance. Our algorithm can still generate adversarial noise in this scenario. However, the resulting average attack success rate 42.12 % is lower than the other two cases.

### 4.3. Adversarial vulnerability of password

To assess the adversarial vulnerability of passwords<sup>1</sup>, the attack success rate for each password, averaged across all speakers, is computed and displayed in Figure 4. The plot indicates that not all passwords exhibit the same level of vulnerability. Specifically, password 2 (“Okay Google”) is found to be more robust to adversarial attacks, whereas password 5 (“Birthday parties have cupcakes and ice cream”) is observed to be highly vulnerable to such attacks. We aim to investigate the source of this variation in vulnerability across different passwords in future work. Specifically, future work will explore whether the phoneme composition of passwords contributes to their vulnerability against adversarial attacks. If so, the research may identify certain phoneme combinations that create passwords inherently resistant to such type of adversarial attacks.

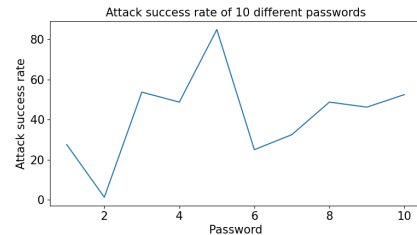


Figure 4: *Attack success rate of 10 different passwords*

## 5. Summary & Future work

This paper highlighted the unrealistic assumption in adversarial attacks against TI-SV system. It proposed a experimental setup for conducting realistic adversarial attacks on the TD-SV system. From the proposed attack algorithm, it demonstrated that the TD-SV system remains vulnerable to realistic adversarial attacks even when the adversarial noise is added at SNR of 30.30 dB. We also examine the vulnerability of 10 different passwords in the evaluation set. The current work focuses on gradient-based white-box attacks, which require detailed information about the TD-SV system to generate adversarial noise. Future research could explore generating adversarial noise using transferable black-box attacks or natural evolution strategies for black-box gradient estimation.

<sup>1</sup>The lexical description of all 10 passwords can be found in the RedDots dataset manual [23]

## 6. References

- [1] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.
- [2] IDVoice, "Innovative voice verification software from id r&d." [Online]. Available: <https://www.idrnd.ai/voice-biometrics/>
- [3] T. B. voiceprint. [Online]. Available: <https://www.tdbank.com/bank/tdvoiceprint.html>.
- [4] ASVSpooF, "Automatic speaker verification and spoofing countermeasures challenge." [Online]. Available: <https://www.asvspoof.org/>
- [5] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–5.
- [6] V. Shchemelinin and K. Simonchik, "Examining vulnerability of voice verification systems to spoofing attacks by means of a tts system," in *Speech and Computer: 15th International Conference, SPECOM 2013, Pilsen, Czech Republic, September 1-5, 2013. Proceedings 15*. Springer, 2013, pp. 132–137.
- [7] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 1962–1966.
- [8] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on gmm i-vector based speaker verification systems," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6579–6583.
- [9] J. Villalba, Y. Zhang, and N. Dehak, "x-vectors meet adversarial attacks: Benchmarking adversarial robustness in speaker verification," in *INTERSPEECH, 2020*, pp. 4233–4237.
- [10] 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.
- [11] W. Zhang, S. Zhao, L. Liu, J. Li, X. Cheng, T. F. Zheng, and X. Hu, "Attack on practical speaker verification system using universal adversarial perturbations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2575–2579.
- [12] S. Sankala, S. R. M. Kodukula *et al.*, "Signal processing interpretation for adversarial examples in speaker verification," in *2024 National Conference on Communications (NCC)*. IEEE, 2024, pp. 1–6.
- [13] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 694–711.
- [14] S. Sankala and S. R. M. Kodukula, "On adversarial vulnerability of activation functions in automatic speaker verification system," *Procedia Computer Science*, vol. 222, pp. 613–623, 2023.
- [15] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2016.
- [16] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical adversarial attacks against speaker recognition systems," in *Proceedings of the 21st international workshop on mobile computing systems and applications*, 2020, pp. 9–14.
- [17] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, and B. Yuan, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 1738–1742.
- [18] Q. Wang, P. Guo, and L. Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," *arXiv preprint arXiv:2005.10637*, 2020.
- [19] S. Sreekanth and K. Sri Rama Murty, "Defending adversarial attacks against asv systems using spectral masking," *Circuits, Systems, and Signal Processing*, pp. 1–21, 2024.
- [20] H. Zeinali, H. Sameti, and L. Burget, "Hmm-based phrase-independent i-vector extractor for text-dependent speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1421–1435, 2017.
- [21] Z. Chen and Y. Lin, "Improving x-vector and plda for text-dependent speaker verification," in *INTERSPEECH, 2020*, pp. 726–730.
- [22] B. Han, Z. Chen, Z. Zhou, and Y. Qian, "The sjt system for short-duration speaker verification challenge 2021," *arXiv preprint arXiv:2208.01933*, 2022.
- [23] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brümmer, D. Van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma *et al.*, "The reddots data collection for speaker recognition," in *Interspeech 2015*, 2015.
- [24] A. Larcher, K. A. Lee, B. Ma, and H. Li, "The rsr2015: Database for text-dependent speaker verification using multiple pass-phrases," in *Annual Conference of the International Speech Communication Association (Interspeech), 2012*.
- [25] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.
- [26] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.
- [27] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [28] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [29] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang, "Unispeech: Unified speech representation learning with labeled and unlabeled data," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 937–10 947.
- [30] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [31] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.
- [32] J. Peng, O. Plchot, T. Stafylakis, L. Mošner, L. Burget, and J. Černocký, "An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 555–562.
- [33] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Text-dependent speaker verification (tdsv) challenge 2024: Challenge evaluation plan." *arXiv preprint arXiv:1xxx.0xxxx*, Tech. Rep., 2024.
- [34] S. Sreekanth, "Exploring self-supervised representations for text-dependent speaker verification," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 1232–1239.
- [35] H. Zeinali, H. Sameti, L. Burget, J. Černocký, N. Maghsoodi, and P. Matejka, "i-vector/hmm based text-dependent speaker verification system for reddots challenge." in *InterSpeech*, 2016, pp. 440–444.
- [36] D. Diatlova, A. Udalov, V. Shutov, and E. Spirin, "Adapting wavlm for speech emotion recognition," *arXiv preprint arXiv:2405.04485*, 2024.