

# Voice Morphing: Two Identities in One Voice

Sushanta K. Pani, Anurag Chowdhury, Morgan Sandler, Arun Ross  
Michigan State University, USA  
rossarun@cse.msu.edu

**Abstract**—In a biometric system, each biometric sample or template is typically associated with a single identity. However, recent research has demonstrated the possibility of generating “morph” biometric samples that can successfully match more than a single identity. Morph attacks are now recognized as a potential security threat to biometric systems. However, most morph attacks have been studied on biometric modalities operating in the image domain, such as face, fingerprint, and iris. In this preliminary work, we introduce Voice Identity Morphing (VIM) - a voice-based morph attack that can synthesize speech samples that impersonate the voice characteristics of a pair of individuals. Our experiments evaluate the vulnerabilities of two popular speaker recognition systems, ECAPA-TDNN and x-vector, to VIM, with a success rate (MMPMR) of over 80% at a false match rate of 1% on the Librispeech dataset.

**Index Terms**—Identity Morphing, Morph Attack, Speaker Recognition, Speech Synthesis

## I. INTRODUCTION

Biometric systems use physical or behavioral traits to recognize individuals [1]. A biometric system acquires a biometric sample of an individual (e.g., voice) using a sensor (e.g., microphone) and extracts a salient feature set (or template). This template is then used to recognize the individual. Typically, a template is associated with a single identity. However, over the past decade, several adversarial techniques, called *morph attacks*, have been developed to create synthetic biometric samples that can successfully match multiple identities [2].<sup>1</sup> Furthermore, in recent times, DeepFake based synthetic image generators have been used to launch morph attacks on image-based biometric systems, viz., face, fingerprint, and iris, with high success rates [5]. The success of such attacks can potentially lead to compromise of security in sensitive applications where a single biometric ID card could be shared by two or more individuals for nefarious purposes.

Existing literature on morph attacks demonstrates its potency against biometric modalities such as face, fingerprint, and iris [5], [6], [7]. For example, landmark-based [8], [9] and deep learning-based [10], [11] face morph attacks have been shown to be effective against face recognition systems. Similarly, researchers have shown the possibility of launching a morph attack against iris matchers both at the image level [2], [7] and feature level [12], [13].

The voice modality, on the other hand, has seemingly been spared from morph attacks until now. The use of voice biometrics is especially relevant in some commercial applications, such as digital voice assistants [14] and telephone banking

[15]. The voice morphing attack may be particularly harmful in scenarios where verification of a single identity is essential to proceed. For instance, consider an online spoken language test. In this context, the test-taking system might require the candidate to enroll their voice beforehand to ensure that the same individual appears for the test. This step is typically achieved using a speaker recognition system, designed to prevent an accomplice from taking the test on behalf of the candidate. However, with a voice morphing attack method, the candidate could enroll a morphed combination of their voice and that of an accomplice. This blend would match both identities, allowing the accomplice to take the test on the candidate’s behalf by successfully matching their voice to the enrolled morphed template. This situation, coupled with the rapid adoption of voice biometric-enabled devices and services, has heightened interest in understanding their vulnerabilities to morphing attacks. Therefore, it is essential to investigate the viability and success rate of such attacks on popular speaker recognition systems.

In this paper, we propose a voice morphing technique called Voice Identity Morphing (VIM)<sup>2</sup> that can synthesize artificial voice samples containing the voice characteristics of a pair of identities. Experimentally we show that the morph voice samples generated from two identities can successfully match target audio samples of both constituent identities using two different popular speaker recognition systems. The proposed method uses the DeepTalk network [16] to extract speaker embeddings from two source identities. Then, it performs a feature-level fusion of the two embeddings producing a new embedding corresponding to the morphed identity. Finally, the morphed embedding is input to a Tacotron 2-based Text-to-Speech synthesizer to generate a morphed audio sample.

The main contributions of this preliminary work are as follows: (a) We propose a voice identity morphing technique capable of generating speech samples that can successfully match two identities within the framework of a speaker recognition system. (b) We evaluate and demonstrate the vulnerability of two popular speaker recognition systems, namely x-vector [17] and ECAPA-TDNN [18], to our proposed method. (c) We perform an ablation study to better understand this vulnerability, and we initiate a discussion on potential forensic measures that may counteract it. (d) We propose directions for future study on this topic.

<sup>1</sup>A related vulnerability known as MasterPrint attack [3] or MasterFace attack [4] has also been studied.

<sup>2</sup>Note that *voice morphing* as defined in this work is different from previous use of this terminology in the speech literature, where it denotes modifying an individual’s voice to sound like another individual.

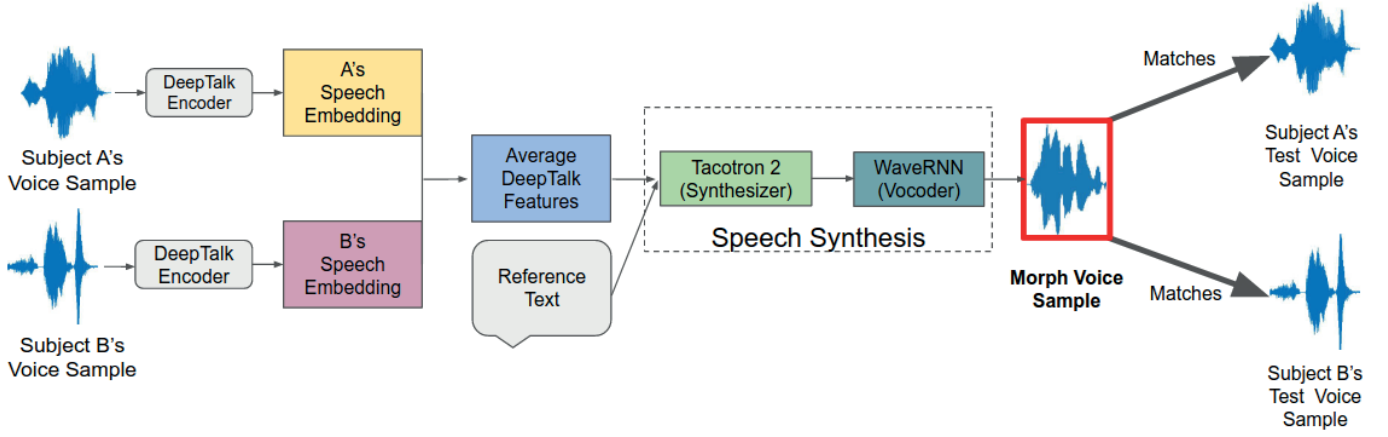


Fig. 1: Illustration of Voice Identity Morphing: Initially, the DeepTalk encoder processes and extracts embeddings that capture the unique speaker characteristics of two distinct individuals. Subsequently, to create a morphed identity, we compute the average of these two embeddings. This averaged embedding then serves as a reference point for our speech synthesis module. Ultimately, employing this reference, the vocoder module generates a spectrogram that merges elements from both contributing speakers.

## II. PROPOSED METHOD: VOICE IDENTITY MORPHING

Voice Identity Morphing (VIM), as shown in Figure 1, has two stages: a) synthetic voice generation and b) morph attack on a speaker recognition system. In the first stage, the proposed method generates synthetic speech samples exhibiting speaker-dependent speech characteristics pertaining to two different speakers, also referred to as the target speaker pair. The synthetic speech sample, called the morphed speech sample, is then compared to individual voice samples from the target speaker pair to launch the morph attack. An attack is successful if the morphed speech sample matches both the target speaker pair's speech samples. The morph voice generation architecture has three separate modules: Encoder, Synthesizer, and Vocoder.

We use a pre-trained **DeepTalk encoder** model to generate vocal style-based speaker identity embeddings of voice samples. We choose this encoder for its competitive performance with the x-vector system and its robustness to degraded audio scenarios. This encoder architecture consists of a 1D-CNN based speech filter bank also known as DeepVOX network [19] and Global Style Token (GST) [20] based prosody embedding network. The DeepVOX network generates short-term speaker-dependent DeepVOX features (see Table I for architecture details). The GST based prosody embedding network generates a fixed dimensional reference embedding from DeepVOX features by using a 2D-CNN followed by a 128-unit GRU. The DeepTalk encoder is pre-trained on the Librispeech, VoxCeleb1 and VoxCeleb2 datasets. The synthesizer module uses these embeddings as an input during the morph sample generation stages (speech synthesis and vocoding). As an initial step of the morph sample generation stage, we average the embeddings ( $Emb_a$  and  $Emb_b$ ) of two voice samples from separate speakers to generate a morph embedding  $Emb_{morph} = (Emb_a + Emb_b)/2$ . We perform this

TABLE I: DeepVOX network setup for learning a 40-dimension feature representation from speech frames. All rows are convolutional layers separated by a SELU activation function.

| In Channels | Out Channels | Kernel | Dilation |
|-------------|--------------|--------|----------|
| 1           | 2            | 5x1    | 2x1      |
| 2           | 4            | 5x1    | 2x1      |
| 4           | 8            | 7x1    | 3x1      |
| 8           | 16           | 9x1    | 4x1      |
| 16          | 32           | 11x1   | 5x1      |
| 32          | 40           | 11x1   | 5x1      |

averaging step to incorporate features from both constituent identities. This assumes that there is an underlying geometric relationship between the identities in the learned embedding space from the DeepTalk encoder. We illustrate these relationships using t-SNE in Figure 3.

We use **Tacotron 2 speech synthesizer** [21] to generate a mel-spectrogram for the corresponding text input. We use the Tacotron 2 synthesizer to retain consistency with the original DeepTalk architecture. Tacotron 2 architecture consists of an encoder and a decoder with an attention mechanism. The encoder creates an internal representation of input text, and the decoder uses the internal representation to generate features that encode the audio as a frame-level mel-spectrogram. The attention mechanism helps the decoder learn from the internal representation by weighting out potential failure cases where some subsequences of text are repeated or ignored by the decoder.

We use a **WaveRNN-based neural vocoder** [22] pretrained model to generate morph samples by inverting the mel-spectrogram output from the Tacotron 2 synthesizer into audio samples. WaveRNN aims to have an expressive and non-linear transformation of the context and minimize the number of operations each step. An RNN addresses this purpose by com-

binning the context and input within a single transformation.

### III. EXPERIMENTAL PROTOCOL

#### A. Dataset

We conducted experiments using the publicly accessible Librispeech dataset [23], an audiobook corpus derived from LibriVox projects. This dataset includes 1000 hours of audio data, in which, for each sample, a speaker reads English text. The dataset is divided into three subsets (100hr, 360hr, 500hr), all sampled at 16kHz. For our experiment, we utilized the 500 hour subset that consists of 1,166 participants (554 female and 612 male). We selected the 500hr subset not only because it is the largest subset, but also because it encompasses 440 speakers, each with more than 30 minutes of speaking time – a factor crucial for the morph generation process.

#### B. Baseline Recognition Performance

We assess speaker recognition systems’ vulnerability to morph samples using two popular speaker recognition systems: x-vector [17] and ECAPA-TDNN [18]. We choose these systems as they are freely available and are used in a wide range of systems.<sup>3</sup> We use the implementation of these systems in Speechbrain [25] toolkit. The x-vector matcher is a TDNN (Time delay neural network) architecture and applies statistical pooling to extract 512-dimensional embedding for variable length utterances. The matcher utilizes categorical cross-entropy loss for training. The ECAPA-TDNN matcher architecture consists of convolutional layers, residual blocks, and attentive statistical pooling layers. It utilizes Additive Margin SoftMax Loss to generate a 192-dimensional embedding. Both matchers utilize Voxceleb1 [26] and Voxceleb2 [27] datasets to train the models. They use cosine distance similarity of speaker embeddings to compare a pair of speaker identities.

Before assessing their vulnerability, we evaluate the baseline recognition performance of these speaker recognition systems on 440 subjects in the 500-hr subset of the Librispeech dataset [23]. Table II provides the performance of these speaker recognition systems in terms of True Match Rate (TMR) at 1%, 0.1%, and 0.01% False Match Rate (FMR). TMR is the proportion of genuine samples that were correctly matched, whereas FMR was the proportion of impostor samples that were incorrectly matched. ECAPA-TDNN model performs better than the x-vector model in correctly classifying genuine and impostor pairs.

#### C. Morph Generation Setup and Results

To generate morph voice samples that incorporate both identities of two different speakers, we first fine-tune a separate Tacotron 2 synthesizer with speech samples of that speaker pair. A pre-trained Tacotron 2 synthesizer needs approximately 30 minutes of the voice samples for fine-tuning [16]. Therefore, we select 440 speakers (221 female and 219 male) which has 30 minutes or more cumulative duration of voice samples.

<sup>3</sup>ECAPA-TDNN amassed 553,704 downloads in one month (June 2023) according to the HuggingFace website [24]

TABLE II: Performance of two speaker recognition systems in terms of TMR (%) at 1%, 0.1%, and 0.01% FMR in the Librispeech dataset. The ECAPA-TDNN and x-vector are two popular, high-performing speaker recognition systems available in the Speechbrain toolkit.

| Matcher    | TMR (%) |          |           |
|------------|---------|----------|-----------|
|            | FMR 1%  | FMR 0.1% | FMR 0.01% |
| ECAPA-TDNN | 98.91   | 97.50    | 93.25     |
| x-vector   | 88.17   | 78.57    | 68.52     |

From 440 speakers, we generate 96,580 speaker pairs ( $^{440}C_2$ ). To generate *better quality* morph samples, we consider those speaker pairs which have *high similarity* in their speech. Each instance of Tacotron 2 takes 8-10 hours to fine-tune. Given this, we select the top 100 speaker pairs. We measure the similarity by the cosine distance of their ECAPA-TDNN-extracted speaker embeddings. Through this process, we select the top 100 speaker pairs, out of which only 43 pairs have unique speakers. The trimmed list of speaker pairs has 3 cross gender speaker pairs. Considering these 43 speaker pairs, we fine-tune 43 different Tacotron 2 synthesizers in parallel. For fine-tuning these Tacotron 2 synthesizers, we also provide 256-dimensional speaker embeddings extracted from a pre-trained DeepTalk encoder model [16] as input along with a reference text. The fine-tuned Tacotron 2 synthesizer outputs a morphed mel spectrogram which is then fed as input into the WaveRNN vocoder [22] to generate morphed speech samples. We create 100 such morphed samples from each speaker pair (10 samples per speaker) which results in 4,300 morphed samples. The speech samples used to generate morph samples are different from the ones used for training the Tacotron 2 synthesizer. We use the remaining voice samples of a speaker for testing. Our experiment has disjoint sets of training (60%), morph (10%) and test (30%) speech samples.

To evaluate the vulnerabilities of the two speaker recognition systems against the generated morph samples (morph attack), we use the Mated Morph Presentation Match Rate (MMPMR) [6] and Morphing Attack Potential (MAP) [28] metrics. MMPMR is a fraction of successful morph attacks out of the total number of morph attacks. A morph attack is considered successful when the morph sample matches with test samples of both speakers. Table III provides the performance of morph attacks in terms of MMPMR at different thresholds corresponding to 1%, 0.1%, and 0.01% FMRs. We report the morph attack success rate in two categories: speaker pair level and morph sample level. A successful morph attack at the speaker pair level has at least one morph sample that matches the samples of both speakers. However, morph sample-level MMPMR reports the success of all morph samples irrespective of the speaker. The proposed morphing technique VIM can create morph samples attacking ECAPA-TDNN and x-vector speaker recognition systems with 95.34% and 86.04% respective success rates at 0.1% FMR, for speaker pair level. The results show that the ECAPA-TDNN speaker recognition

system is more susceptible to morph attacks compared to the x-vector recognition system. The considerable success rate of morph attacks could likely be related to the morph pair selection process or the effective capturing of subject information by the DeepTalk encoding method. This infers that prior knowledge of the speaker recognition system would generate stronger morph attacks. Also, we hypothesize that state-of-the-art speaker recognition systems are likely to detect vocal features of both the parent speakers in a composite audio. This may make them vulnerable to such morphing attacks as well. We find that the fusion of speech synthesis embeddings generates effective morph audio samples for use in attacks on speaker recognition systems.

TABLE III: Vulnerability assessment of two speaker recognition systems to voice identity morph attack in terms of MMPMR (%) at different threshold corresponding to 1%, 0.1%, and 0.01% FMR on the Librispeech dataset.

| Matcher    | Speaker pair MMPMR (%) |          |           | Morph sample MMPMR (%) |          |           |
|------------|------------------------|----------|-----------|------------------------|----------|-----------|
|            | FMR 1%                 | FMR 0.1% | FMR 0.01% | FMR 1%                 | FMR 0.1% | FMR 0.01% |
| ECAPA-TDNN | 100.00                 | 95.34    | 81.39     | 91.23                  | 62.11    | 21.58     |
| x-vector   | 93.02                  | 86.04    | 9.30      | 82.13                  | 38.95    | 4.32      |

#### D. Result Analysis

We further analyze our morph attack performance using: 1) histogram plots, 2) t-SNE plots, and 3) morphing attack potential (MAP). Figure 2 shows the histogram plots of match scores corresponding to genuine pairs (green), impostor pairs (red), and pairs which include at least one morphed sample (blue) for both speaker recognition systems. In both systems, we find that the morphed pairs match score distribution lies between genuine and impostor score distributions. Morph samples are classified as genuine matches in the ECAPA-TDNN and x-vector systems with recognition thresholds of 0.46 and 0.96 respectively at 0.1% FMR.

The second analysis we perform is based on the t-SNE dimensionality reduction technique. The t-SNE [29] method helps visualize high-dimensional embeddings in a two-dimensional space by reducing the dimension. Figure 3 shows

TABLE IV: Morphing Attack Potential (MAP) [28]: This metric represents the success rate (%) of a morphed sample matching at least a specified number of probe voice samples (denoted as # of attempts) within the Librispeech dataset, using one or both of the speaker recognition systems (SRS), namely ECAPA-TDNN and x-vector. The success rate is evaluated at three false match rate (FMR) thresholds: 1%, 0.1%, and 0.01%.

| # of Attempts | FMR 1% |       | FMR 0.1% |       | FMR 0.01% |       |
|---------------|--------|-------|----------|-------|-----------|-------|
|               | 1 SRS  | 2 SRS | 1 SRS    | 2 SRS | 1 SRS     | 2 SRS |
| 1             | 92.0%  | 52.7% | 60.4%    | 7.6%  | 20.2%     | 2.3%  |
| 2             | 90.2%  | 46.3% | 54.4%    | 5.7%  | 16.5%     | 1.7%  |
| 3             | 88.9%  | 41.6% | 50.8%    | 5.0%  | 14.3%     | 1.0%  |
| 4             | 87.9%  | 38.1% | 47.9%    | 4.6%  | 13.0%     | 0.6%  |
| 5             | 87.0%  | 35.7% | 45.8%    | 4.2%  | 11.4%     | 0.3%  |

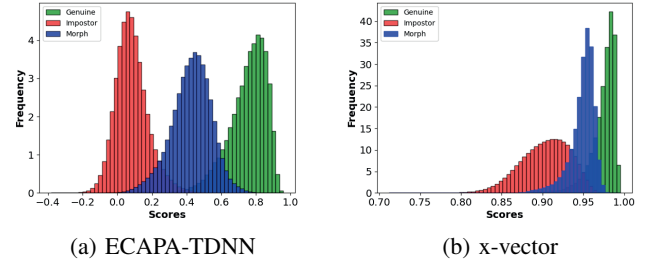


Fig. 2: Speaker recognition match score distributions of non-morph versus non-morph genuine (Green), non-morph versus non-morph impostor (Red), and morph versus non-morph genuine morph scores (Blue) using ECAPA-TDNN and x-vector embeddings.

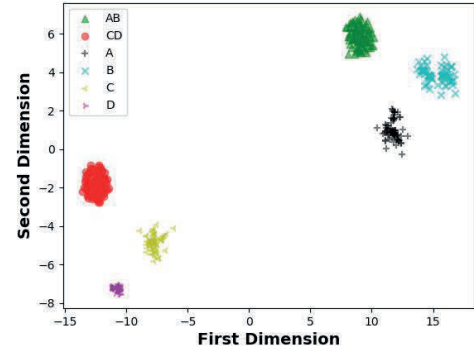


Fig. 3: t-SNE plot which illustrates high-dimensional ECAPA-TDNN embeddings of morph speech samples from two separate speaker pairs (AB and CD) and non-morph speech samples of individual speakers (A, B, C, and D). Morph embeddings of each pair are closer to the non-morph embeddings of their constituent speakers.

the t-SNE plot of morph sample embeddings from two speaker pairs (AB and CD) along with non-morph samples of four constituent speakers (A, B, C, and D). The embeddings are extracted by the ECAPA-TDNN recognition system. Here, embeddings of morph samples of one speaker pair (AB) are closer to embeddings of A and B speakers. Similarly, embeddings of morph samples of another speaker pair (CD) are closer to embeddings of C and D speakers. The analysis again validates the effectiveness of the proposed morphing technique and the potential threat of morph attacks.

The Morphing Attack Potential (MAP) [28] constitutes the third analysis. This metric takes into account multiple Speaker Recognition Systems (SRS) to ensure generality, and a variable number of verified probe samples for robustness. The result is a matrix (Table IV) in which one axis represents the number of probe samples (referred to as the number of attempts), and the other axis represents the number of SRS. The entries in each row represent the success rate (in percentage) of a morphed sample matching at least a specified number of probe voice

samples (referred to as the number of attempts) using either or both of the SRS, viz., ECAPA-TDNN and x-vector. We report the success rates over three FMR thresholds of 1%, 0.1%, and 0.01%. The results imply that VIM is effective at a fairly competitive FMR of 1%, but suggest there is still room for improvement in performance at very low FMR thresholds. This may be attributed to the morph selection process or perhaps to the pre-trained models used in the encoder and speech synthesis steps.

#### IV. SUMMARY AND FUTURE WORK

To the best of our knowledge, this preliminary work is the first to demonstrate the vulnerability of speaker recognition systems to morph attacks. In this regard, we propose a voice morphing technique called VIM to generate speech samples corresponding to the identities of two subjects. Using these morph samples, we demonstrate a morph attack success rate of over 80% on two popular speaker recognition systems (ECAPA-TDNN and x-vector). As future work, we propose to select high-similarity pairs for a morphing attack using x-vector to investigate whether the selection process plays a vital role in the performance of such an attack. Additionally, evaluating newer speaker recognition systems such as TitaNet [30] and MFA-Conformer [31] would provide more insight into the generalizability of VIM. Comparing other speech synthesis systems in the speech synthesis step would shed light on the role this step plays in the VIM attack. Furthermore, we aim to develop a system for detecting morphed speech samples, possibly through the identification of their constituent identities. It may also be interesting to explore the maximum number of identities that can be combined into a single audio sample using VIM.

#### V. REPRODUCIBILITY

The code for generating VIM samples can be found online at our Github link.<sup>4</sup>

#### REFERENCES

- [1] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics*. Springer Science & Business Media, 2007.
- [2] M. Ferrara, A. Franco, and D. Maltoni, "The magic passport," in *IEEE International Joint Conference on Biometrics*, 2014, pp. 1–7.
- [3] A. Roy, N. Memon, and A. Ross, "Masterprint: Exploring the vulnerability of partial fingerprint-based authentication systems," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 9, pp. 2013–2025, 2017.
- [4] H. H. Nguyen, S. Marcel, J. Yamagishi, and I. Echizen, "Master face attacks on face recognition systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 398–411, 2022.
- [5] S. Venkatesh, R. Ramachandra, K. Raja, and C. Busch, "Face morphing attack generation and detection: A comprehensive survey," *IEEE Transactions on Technology and Society*, vol. 2, no. 3, pp. 128–145, 2021.
- [6] U. Scherhag, A. Nautsch, C. Rathgeb, and Others, "Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2017, pp. 1–7.
- [7] R. Sharma and A. Ross, "Image-level iris morph attack," in *IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 3013–3017.
- [8] M. Matteo Ferrara, A. Franco, and D. Maltoni, "Decoupling texture blending and shape warping in face morphing," in *Proceedings of the 18th International Conference of the Biometrics Special Interest Group (BIOSIG)*. Gesellschaft für Informatik eV, 2019.
- [9] R. Ramachandra, K. Raja, and C. Busch, "Detecting morphed face images," in *8th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS*, 2016, pp. 1–7.
- [10] H. Zhang, S. Venkatesh, R. Ramachandra, and Others, "MIP-GAN—generating strong and high quality morphing attacks using identity prior driven GAN," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 365–383, 2021.
- [11] N. Damer, A. M. Saladie, A. Braun, and A. Kuijper, "MorGAN: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network," in *IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*, 2018, pp. 1–10.
- [12] M. Ferrara, R. Cappelli, and D. Maltoni, "On the feasibility of creating double-identity fingerprints," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 892–900, 2016.
- [13] C. Rathgeb and C. Busch, "On the feasibility of creating morphed iris-codes," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 152–157.
- [14] M. B. Hoy, "Alexa, siri, cortana, and more: An introduction to voice assistants," *Medical Reference Services Quarterly*, vol. 37, no. 1, pp. 81–88, 2018.
- [15] H. Melin, A. Sandell, and M. Ihse, "CTT-Bank: A speech controlled telephone banking system—an initial evaluation," *TMH-QPSR*, vol. 1, pp. 1–27, 2001.
- [16] A. Chowdhury, A. Ross, and P. David, "DeepTalk: Vocal style encoding for speaker recognition and speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6189–6193.
- [17] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.
- [18] B. Desplanques, J. Thienpondt, and K. Demuyne, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in *INTERSPEECH 2020*, pp. 3830–3834.
- [19] A. Chowdhury and A. Ross, "Deepvox: Discovering features from raw audio for speaker recognition in non-ideal audio signals," *arXiv preprint arXiv:2008.11668*, 2020.
- [20] Y. Wang, D. Stanton, Y. Zhang, and Others, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 5180–5189.
- [21] J. Shen, R. Pang, R. J. Weiss, and Others, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.
- [22] N. Kalchbrenner, E. Elsen, K. Simonyan, and Others, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.
- [23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [24] H. F. Inc., "Hugging face: The ai community building the future," 2023, accessed: 2023-07-08. [Online]. Available: <https://www.huggingface.co>
- [25] M. Ravanelli, T. Parcollet, and Others, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.
- [26] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *INTERSPEECH*, 2017.
- [27] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.
- [28] M. Ferrara, A. Franco, D. Maltoni, and C. Busch, "Morphing attack potential," in *International Workshop on Biometrics and Forensics (IWBIF)*. IEEE, 2022, pp. 1–6.
- [29] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.
- [30] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1D depth-wise separable convolutions and global context," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8102–8106.

<sup>4</sup><https://github.com/morganlee123/VIM>

- [31] Y. Zhang, Z. Lv, H. Wu, and Others, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," in *INTERSPEECH*, 2022.