# De-AntiFake: Rethinking the Protective Perturbations Against Voice Cloning Attacks

Wei Fan, Kejiang Chen, Chang Liu, Weiming Zhang, Nenghai Yu

University of Science and Technology of China

range@mail.ustc.edu.cn chenkj@ustc.edu.cn

中国科学技术大学

University of Science and Technology of China

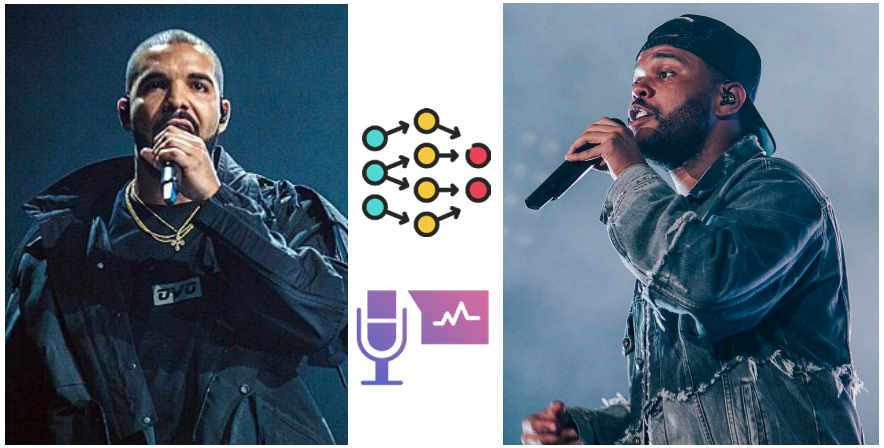
ICML

International Conference On Machine Learning

# Voice Cloning: Useful Apps




Industry leaders utilize advanced AI voice synthesis to create more natural and intelligent Voice AI applications.



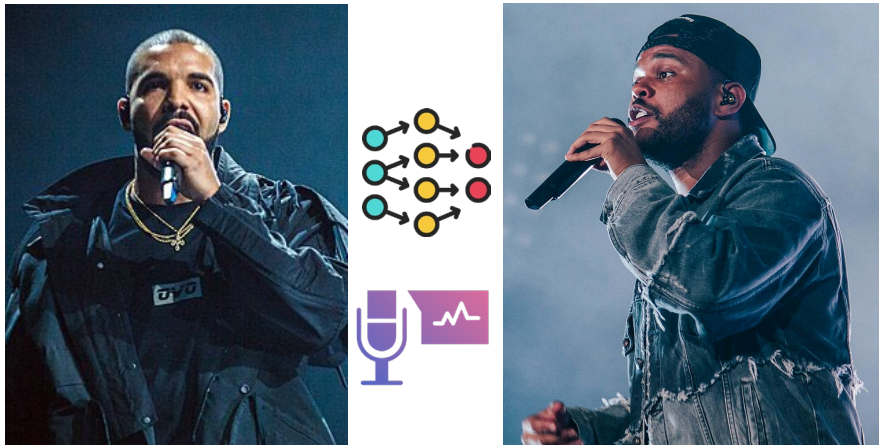
In April 2023, a song using AI to clone the voices of Drake and The Weeknd went viral on social media, garnering over 15 million views on TikTok in just two days.

AI voice synthesis recreated Val Kilmer's voice for his role in *Top Gun: Maverick* after cancer treatment damaged the actor's real voice.

# Voice Cloning: Useful Apps




Industry leaders utilize advanced AI voice synthesis to create more natural and intelligent Voice AI applications.



In April 2023, a song using AI to clone the voices of Drake and The Weeknd went viral on social media, garnering over 15 million views on TikTok in just two days.

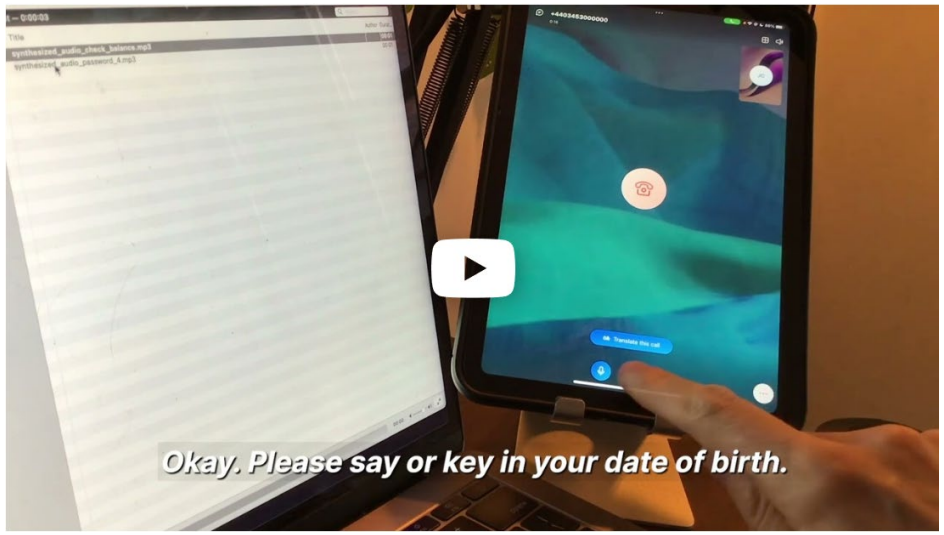
AI voice synthesis recreated Val Kilmer's voice for his role in *Top Gun: Maverick* after cancer treatment damaged the actor's real voice.

Rapid advancements in voice cloning have been widely used in conversational AI, entertainment and accessibility

# Voice Cloning: Security Risks



Finance worker in Hong Kong paid out $25 million after attending a deepfake video call. Scammers cloned the voices and images of senior executives to order fraudulent transfers.

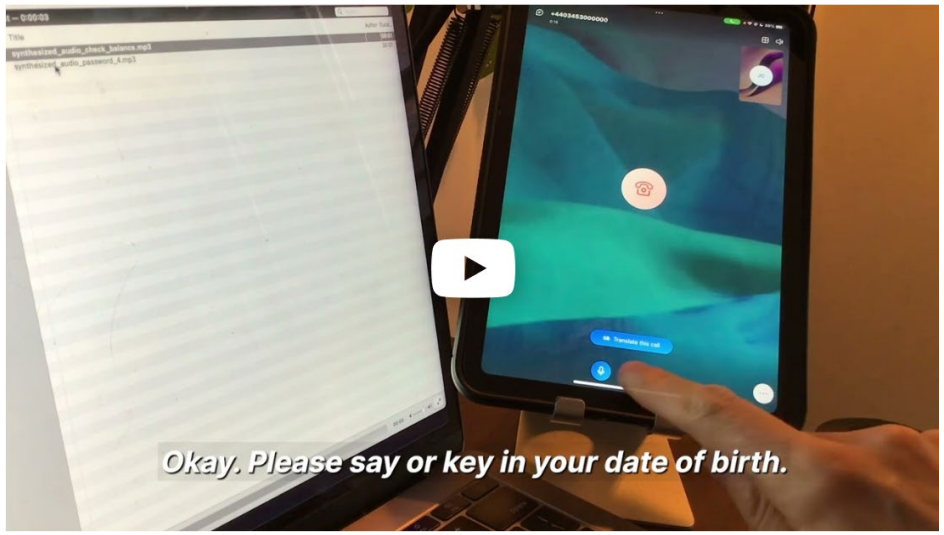



Journalist from *Vice* bypassed a bank's voice verification system using a free AI voice clone, granting him access to his personal account information.

# Voice Cloning: Security Risks



Finance worker in Hong Kong paid out $25 million after attending a deepfake video call. Scammers cloned the voices and images of senior executives to order fraudulent transfers.




Journalist from *Vice* bypassed a bank's voice verification system using a free AI voice clone, granting him access to his personal account information.

Need for effective defenses against malicious voice cloning

# Voice Cloning Attacks



**Victim**

“Good morning”

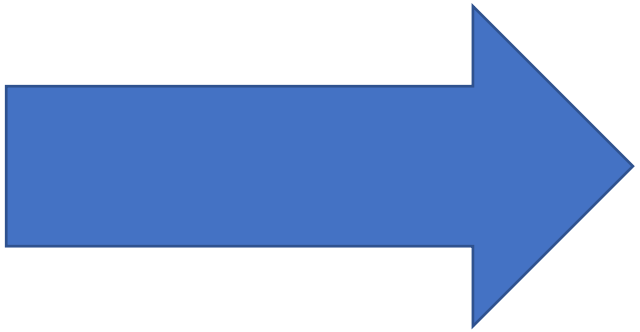
# Voice Cloning Attacks



Victim

“Good morning”

Attacker

Sounds Like Victim

“Send me $50”

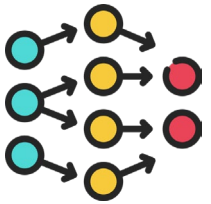
# Voice Cloning Attacks



**Victim**

"Good morning"

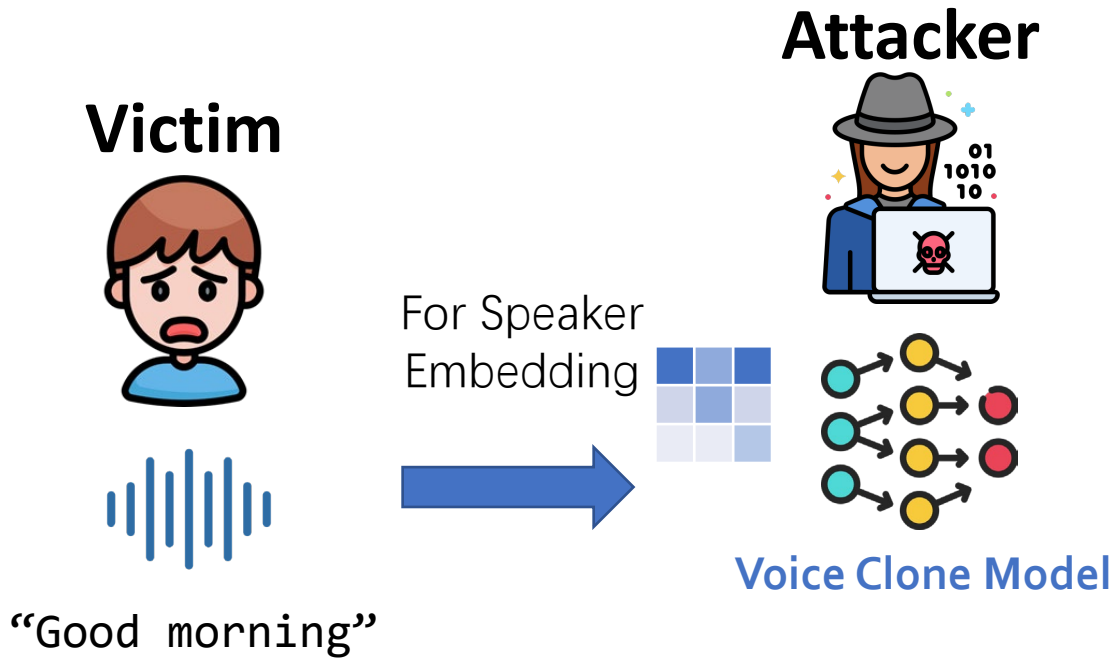
**Attacker**

Voice Clone Model

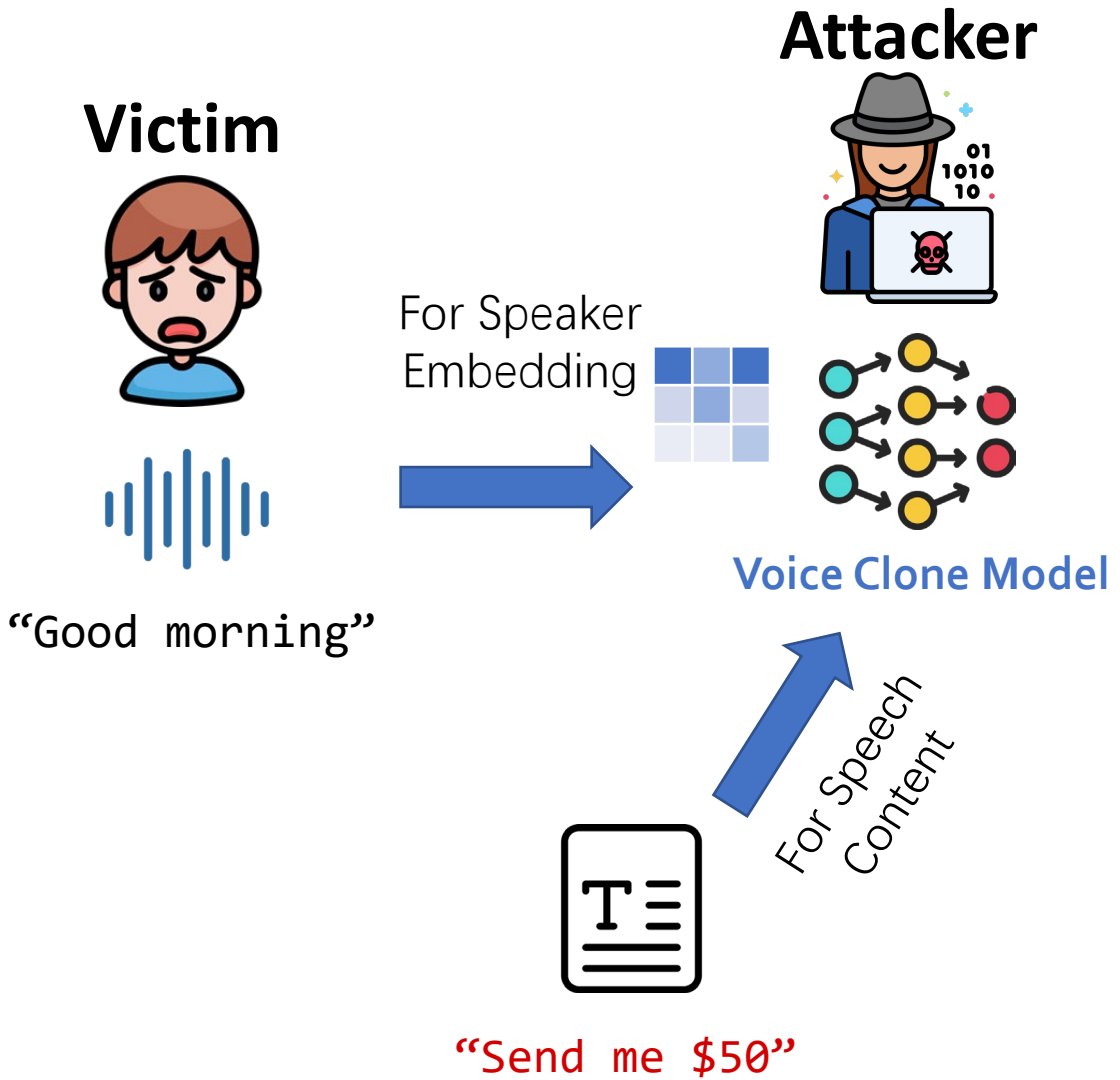
# Voice Cloning Attacks





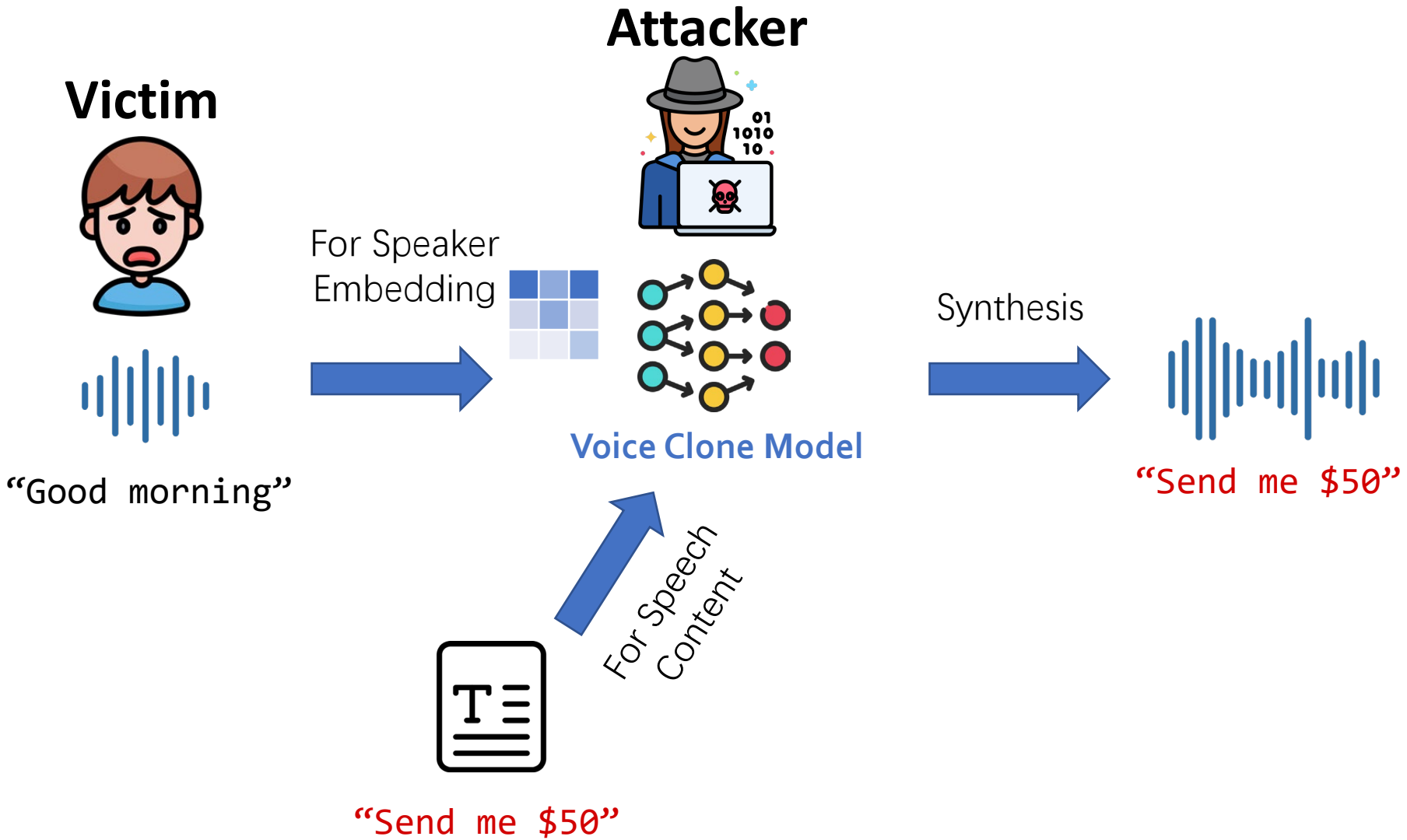
# Voice Cloning Attacks





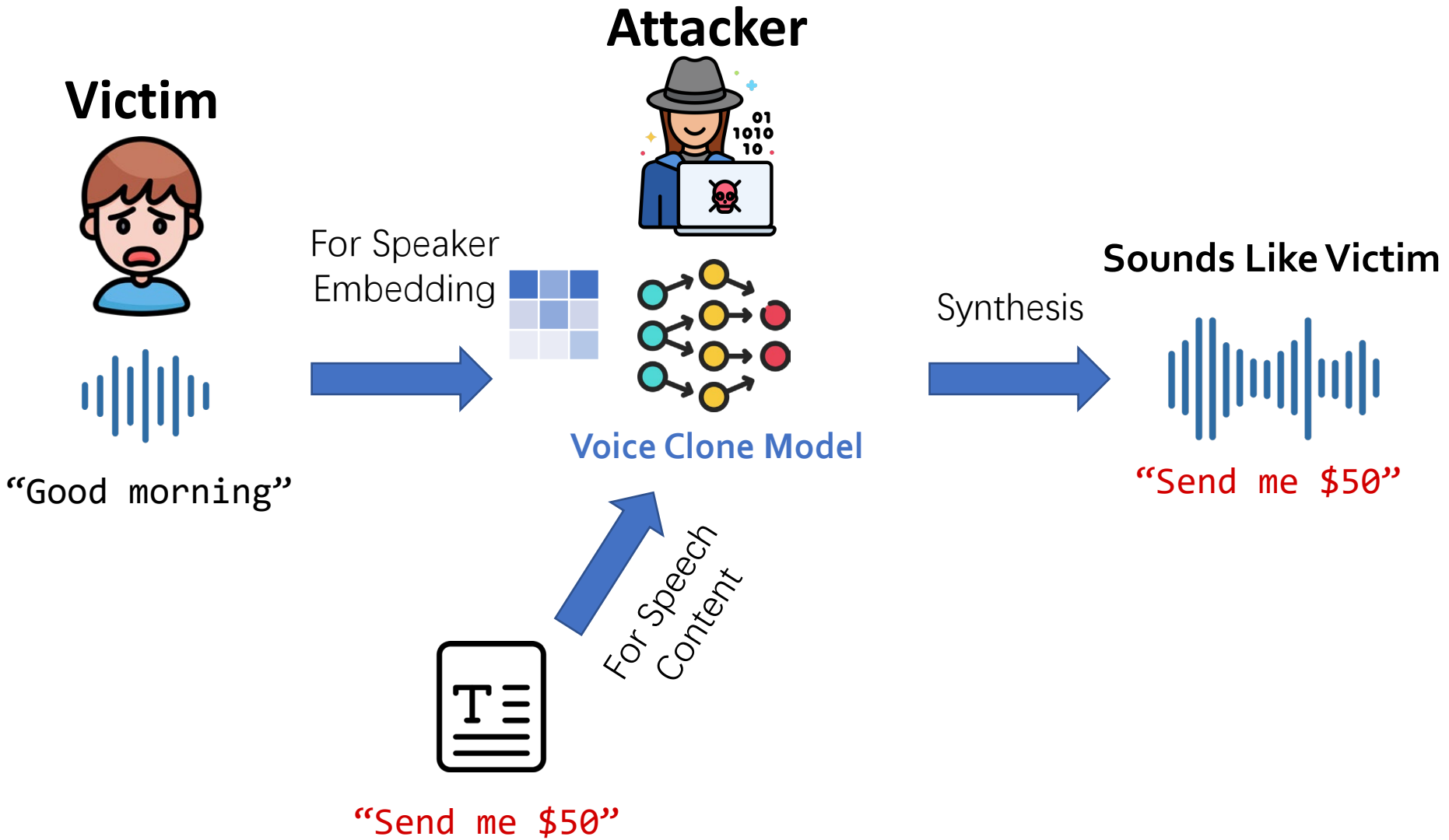
# Voice Cloning Attacks





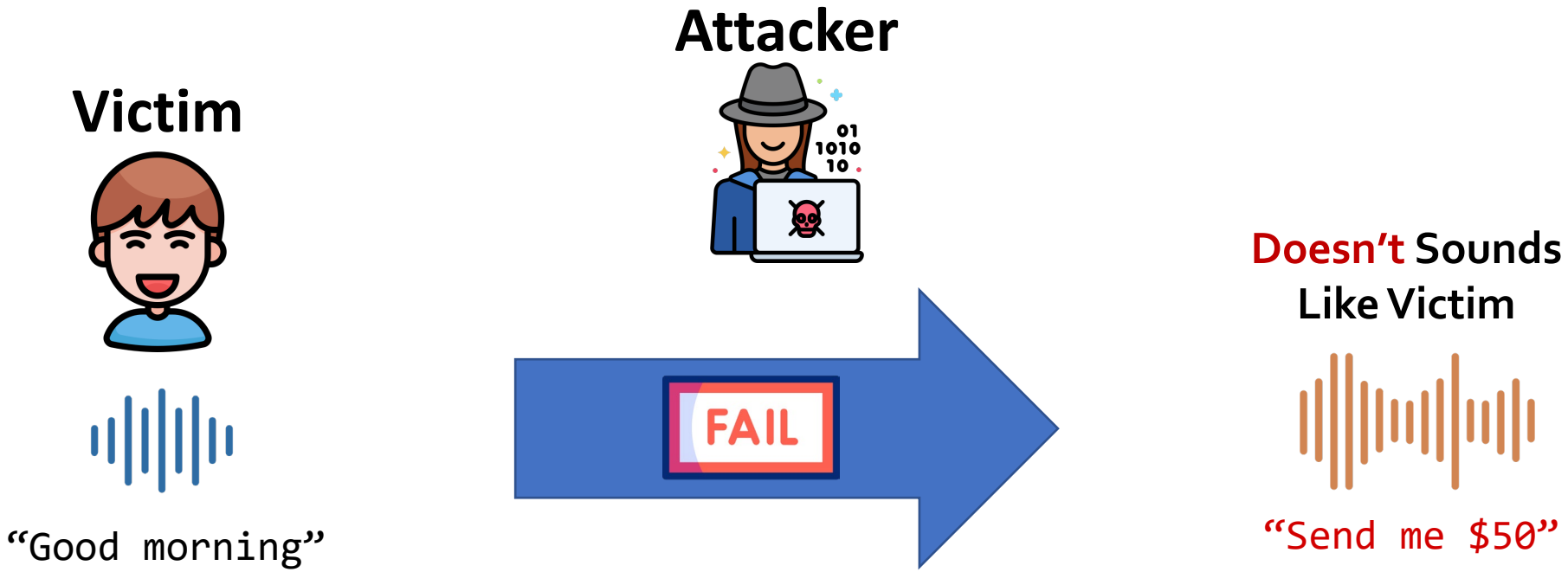
# Voice Cloning Attacks





# Proactive Defense Strategy





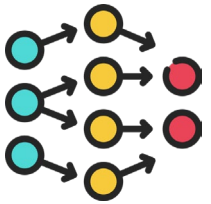
# Existing Defense: Protective Perturbations



**Victim**

“Good morning”

**Attacker**

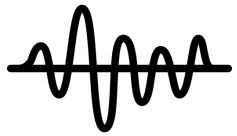



Voice Clone Model

# Existing Defense: Protective Perturbations



Imperceptible Protective Perturbation

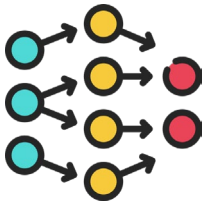


+

Victim

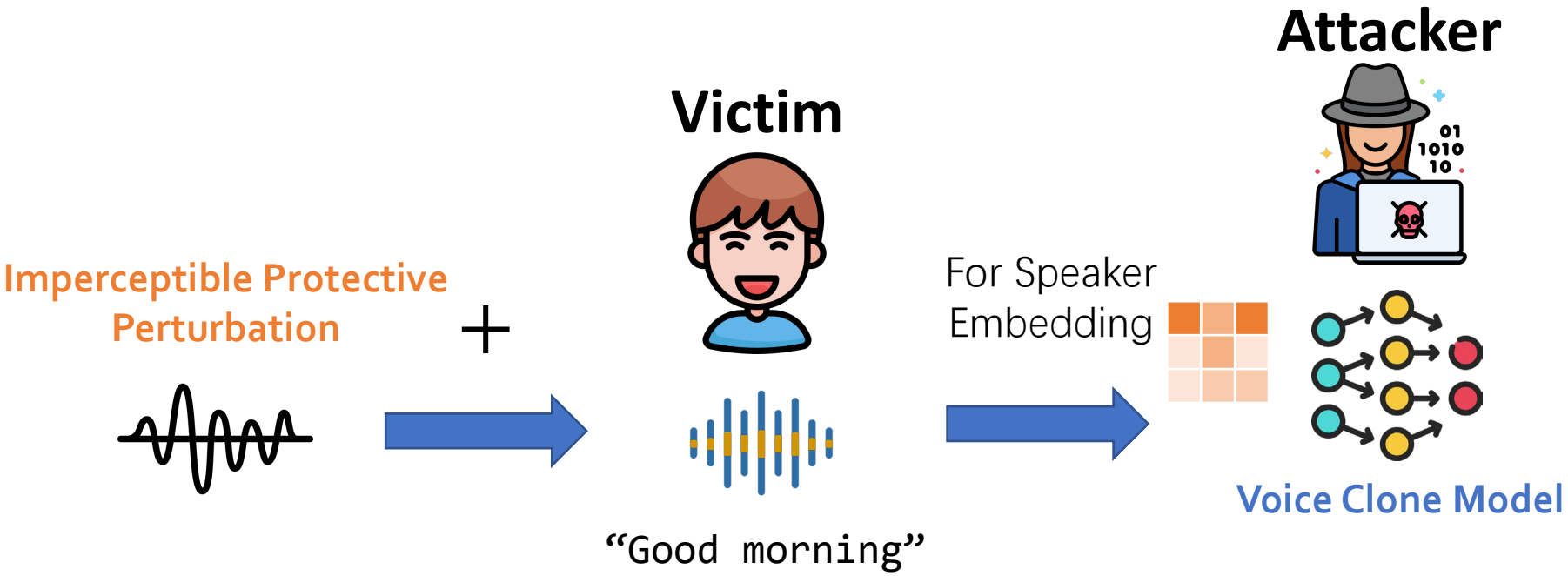
"Good morning"

Attacker

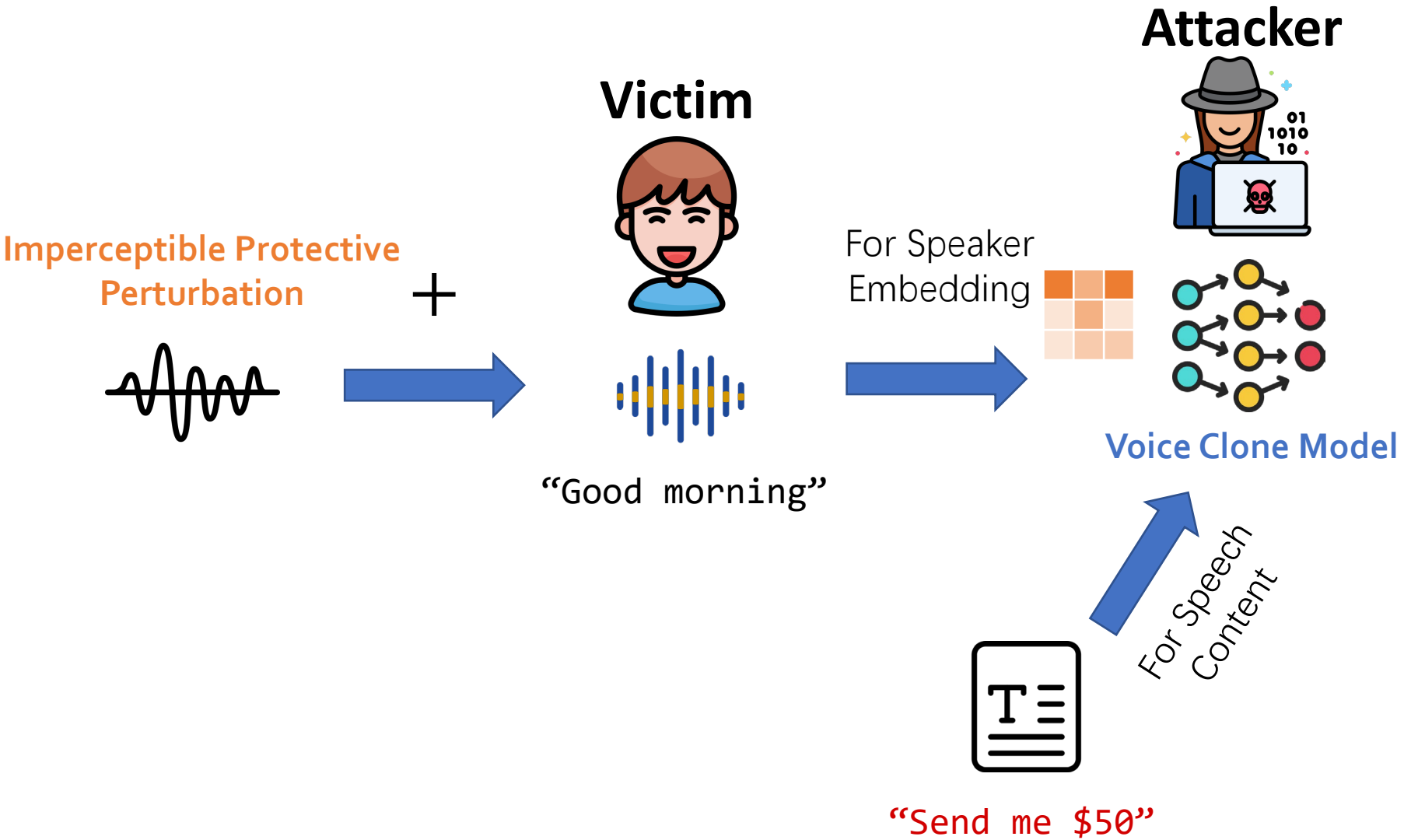
Voice Clone Model

# Existing Defense: Protective Perturbations



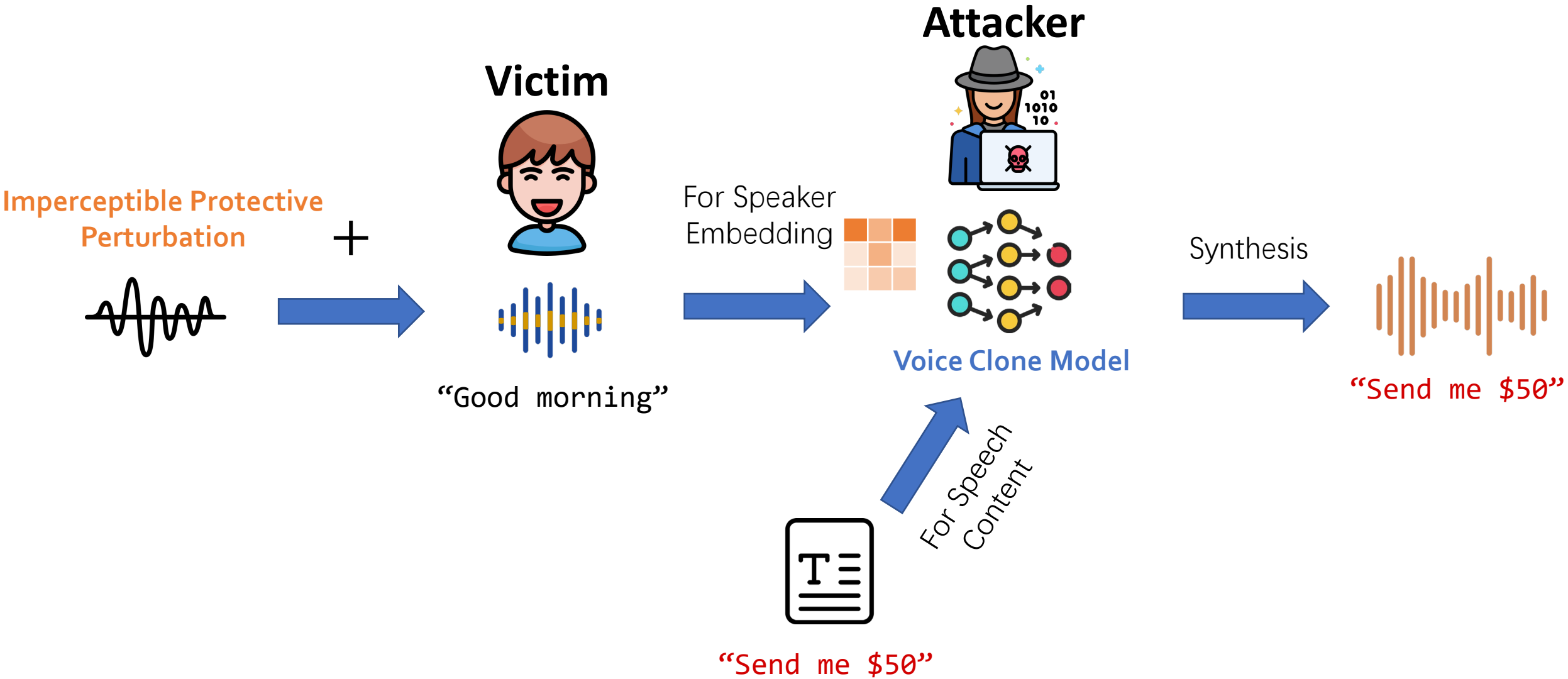
# Existing Defense: Protective Perturbations





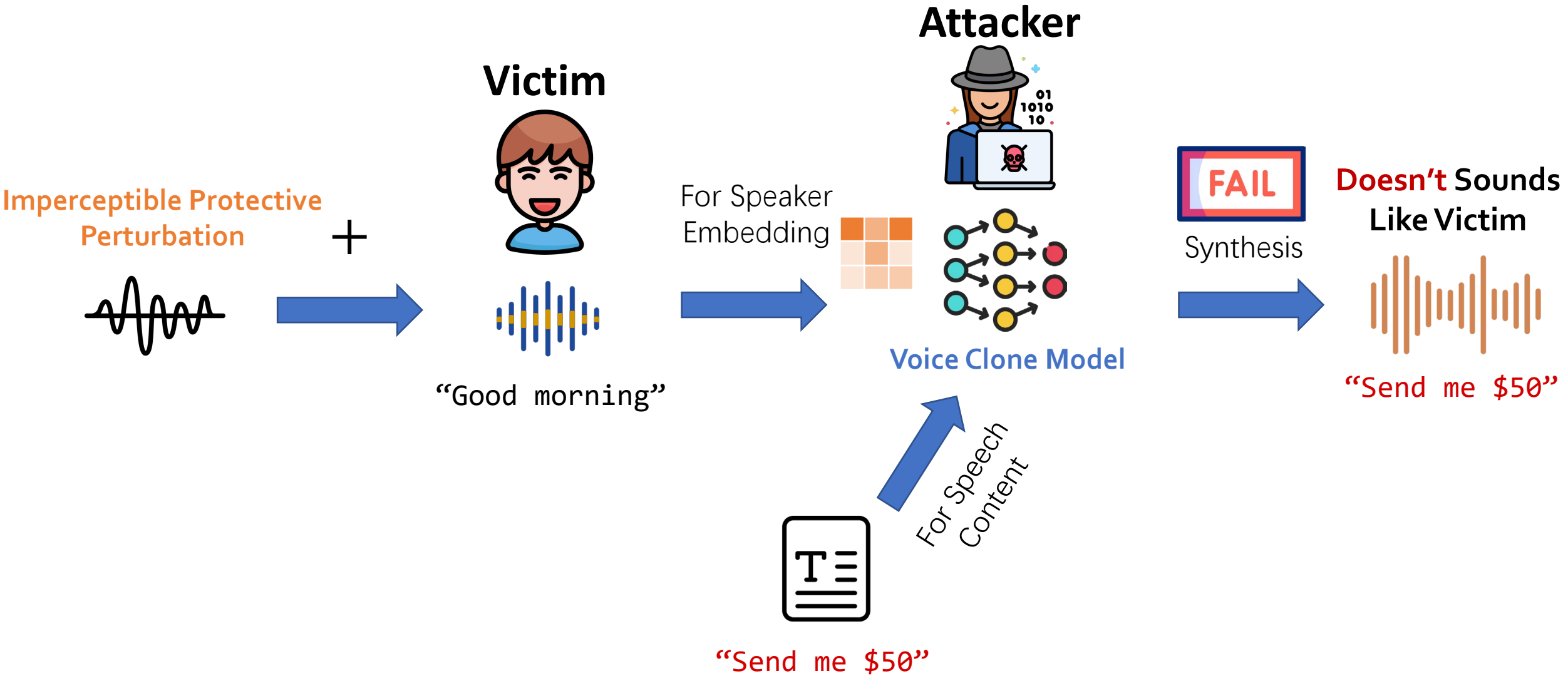
# Existing Defense: Protective Perturbations





# Existing Defense: Protective Perturbations





# Rethinking Current Protective Perturbations



Existing proactive perturbations methods add imperceptible distortions to speech, successfully prevent voice cloning **in ideal conditions**.

# Rethinking Current Protective Perturbations



Existing proactive perturbations methods add imperceptible distortions to speech, successfully prevent voice cloning **in ideal conditions**.

**But what if attackers try to purify these perturbations?**

# Rethinking Current Protective Perturbations



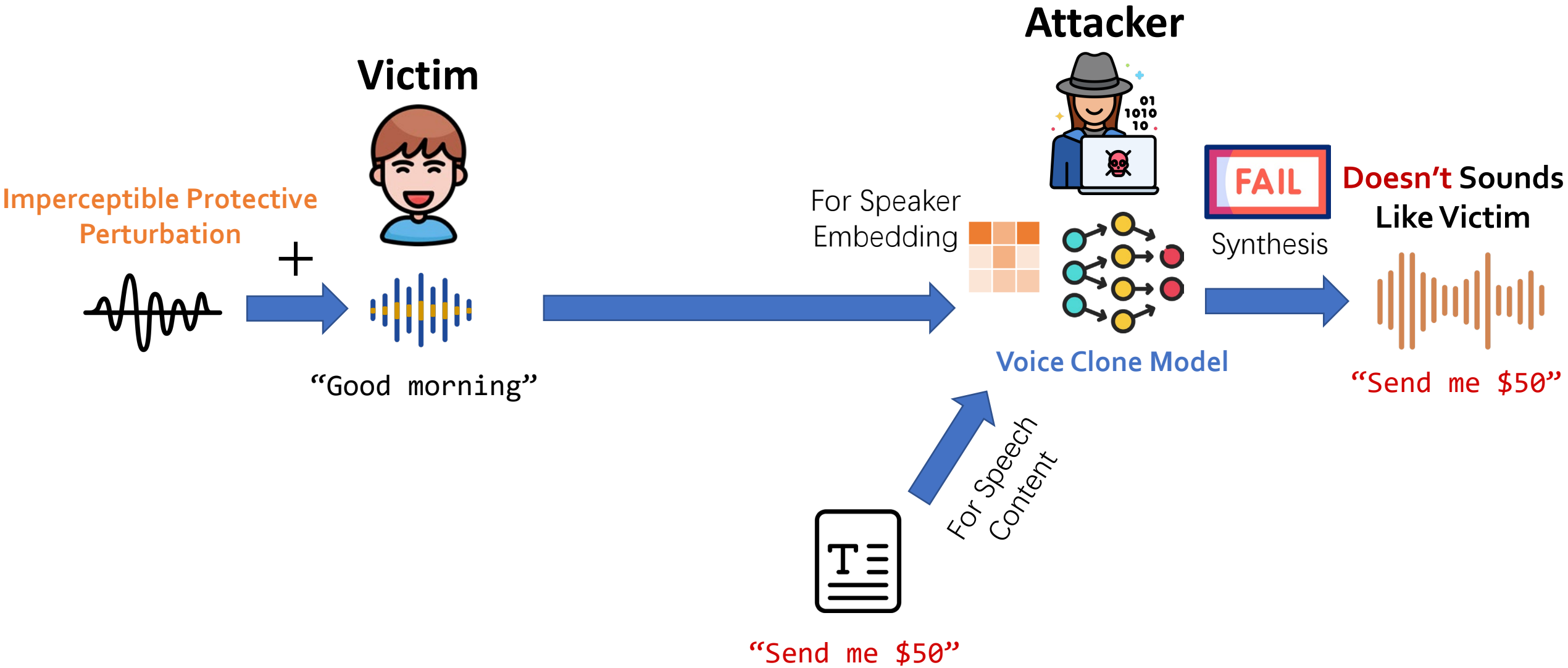
Existing proactive perturbations methods add imperceptible distortions to speech, successfully prevent voice cloning **in ideal conditions**.

**But what if attackers try to purify these perturbations?**

If existing defenses are vulnerable to purification, they **may provide a false sense of security**.

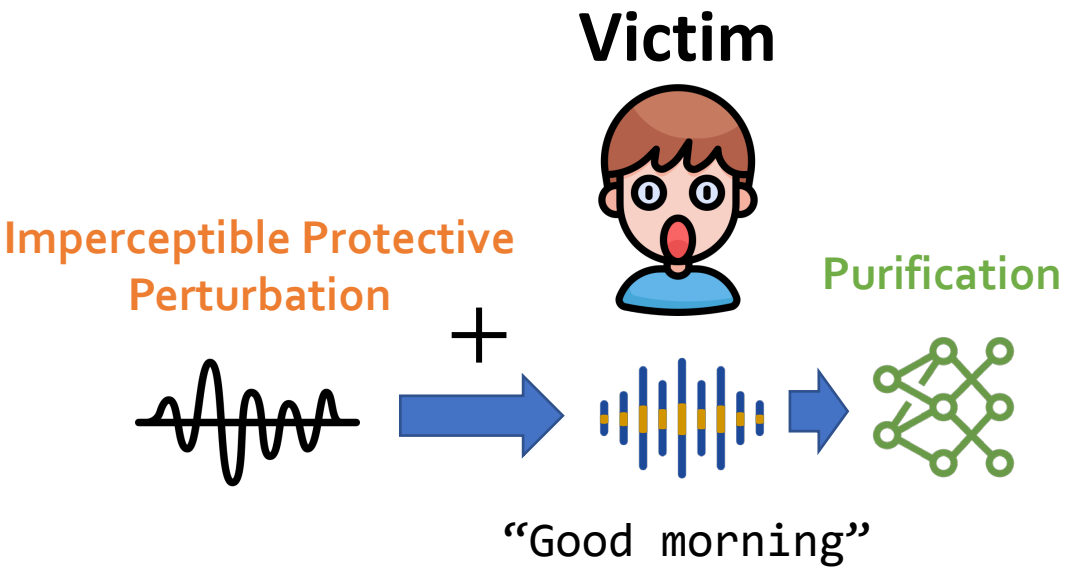
# If the Attackers Try to Purify the Perturbations…



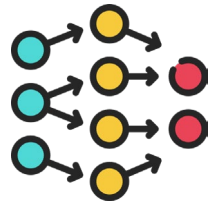


# If the Attackers Try to Purify the Perturbations…





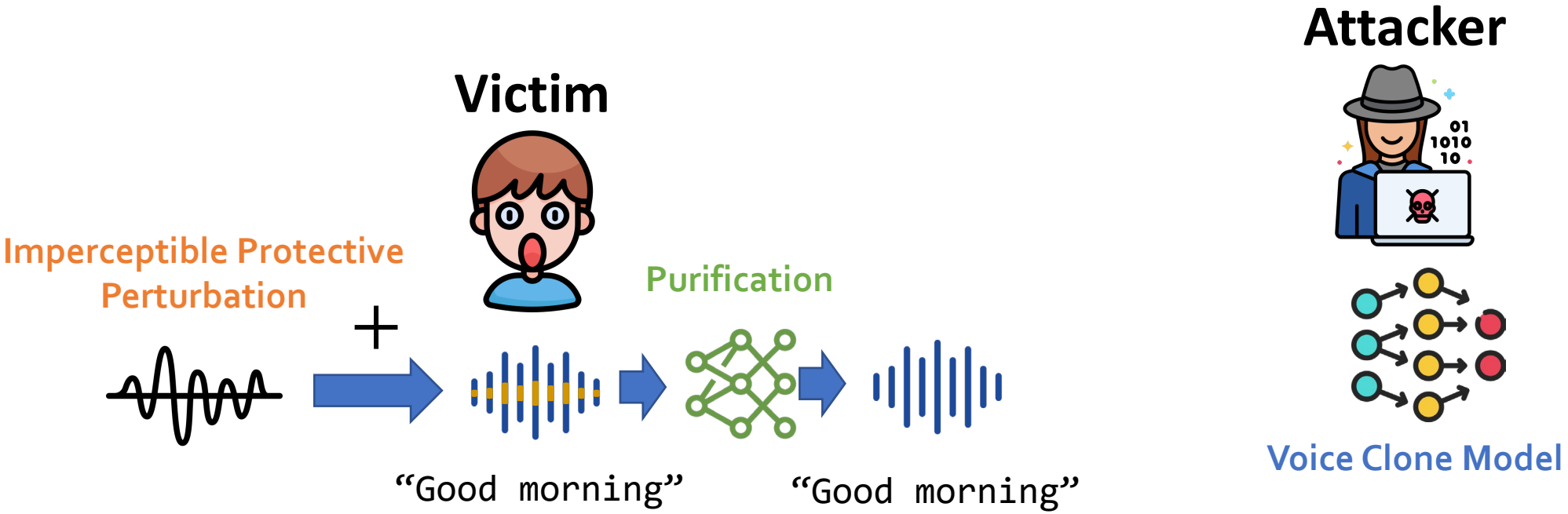




Voice Clone Model

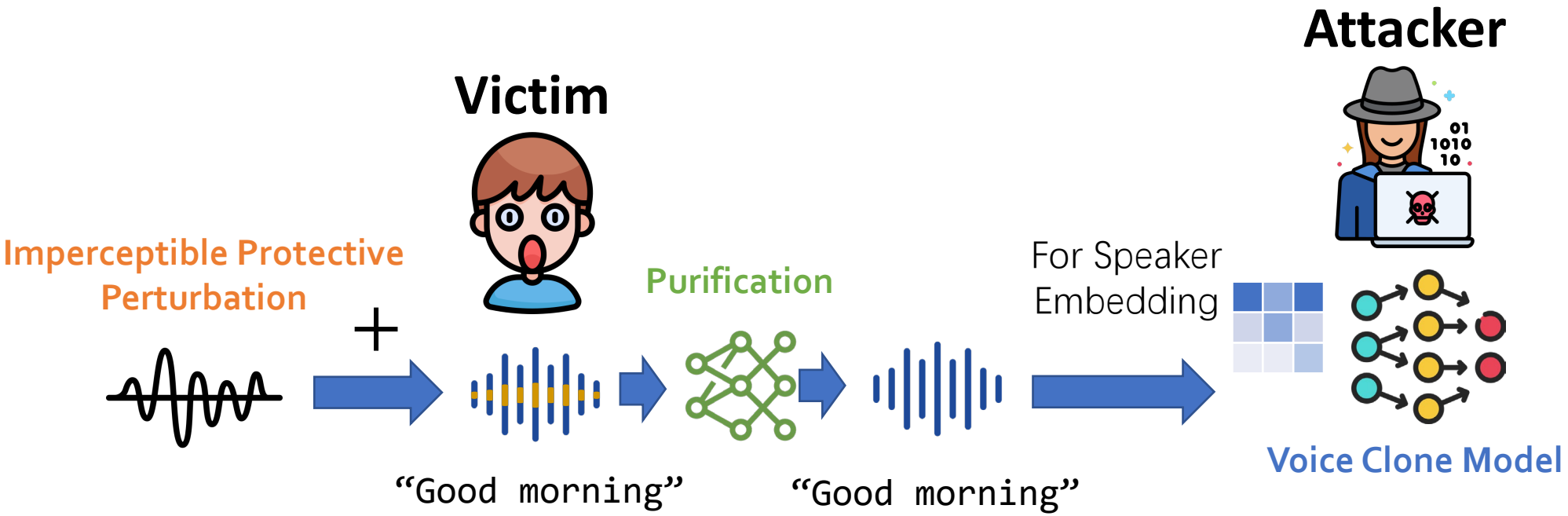
# If the Attackers Try to Purify the Perturbations...





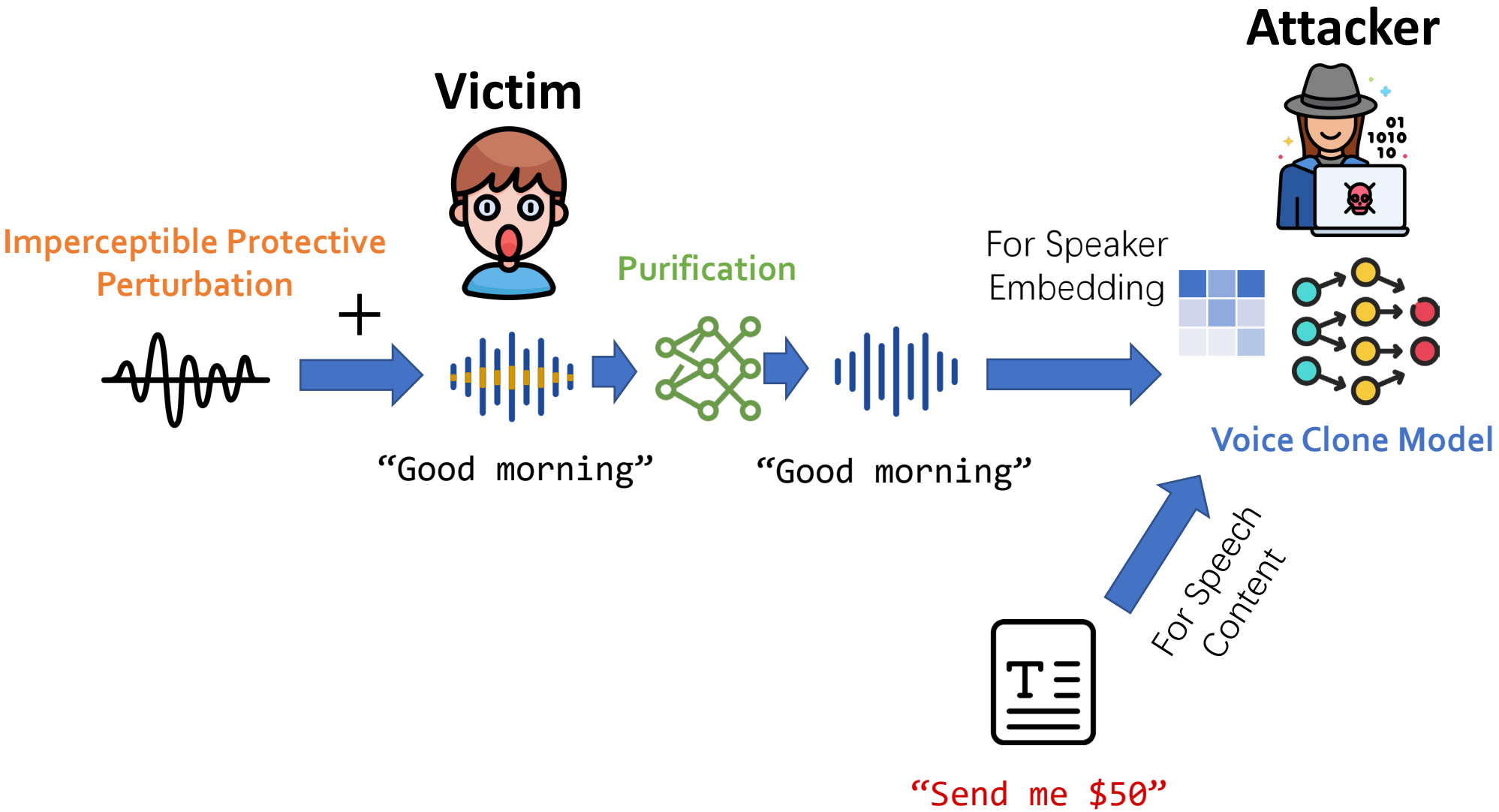
# If the Attackers Try to Purify the Perturbations...





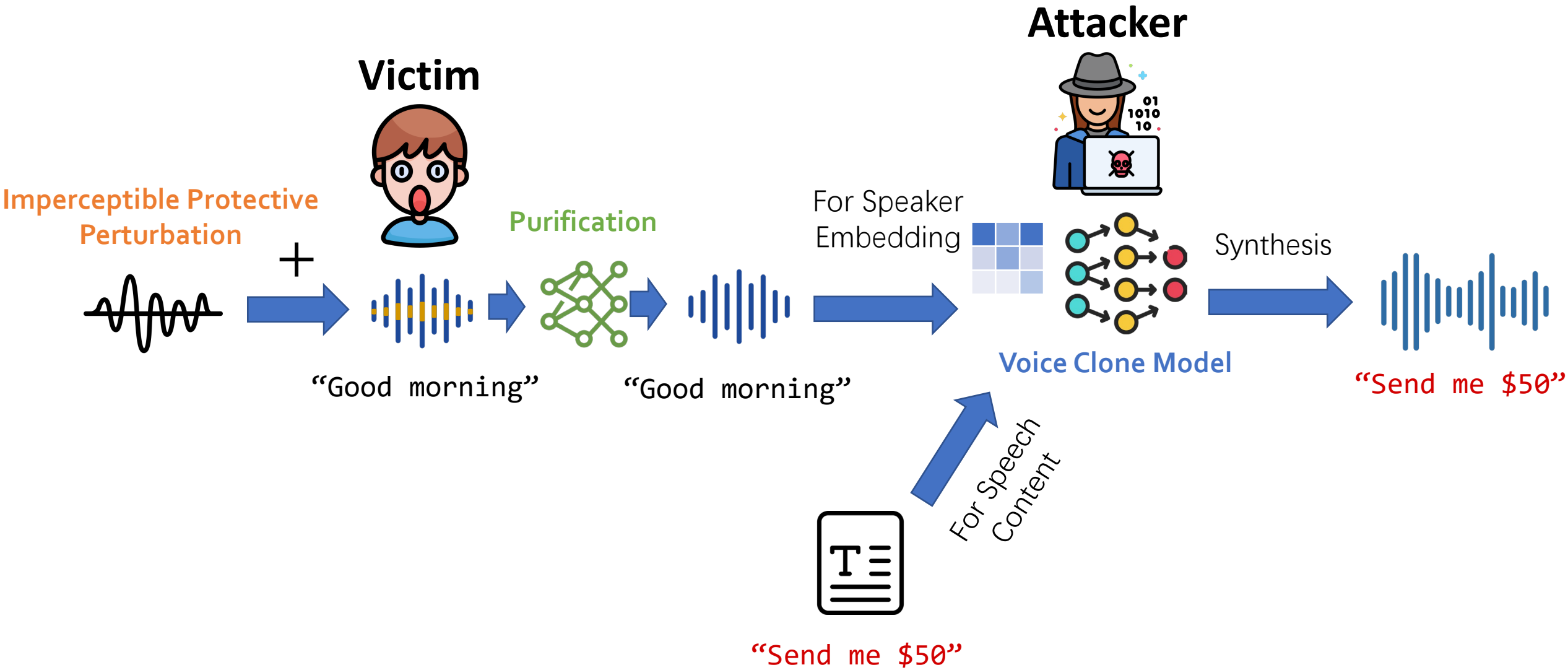
# If the Attackers Try to Purify the Perturbations...





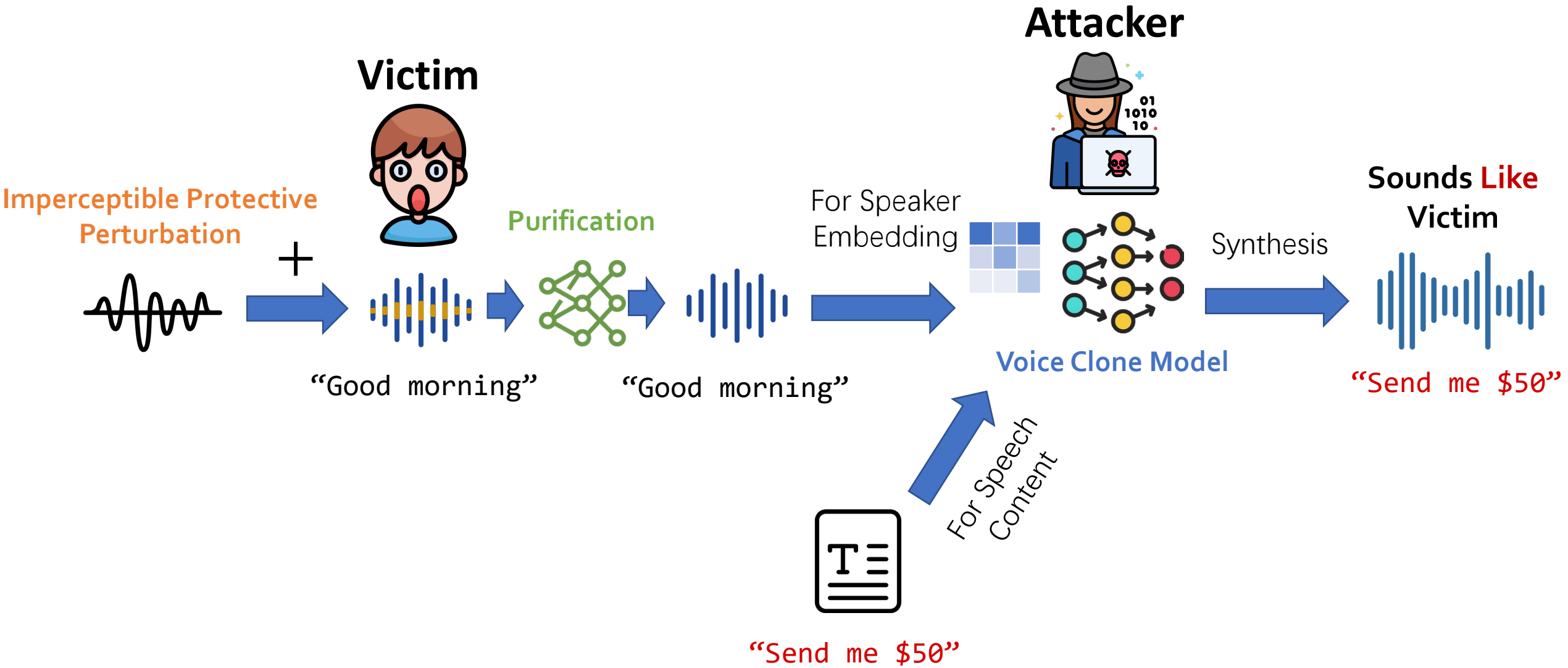
# If the Attackers Try to Purify the Perturbations...





# If the Attackers Try to Purify the Perturbations…





# Our Contribution 1



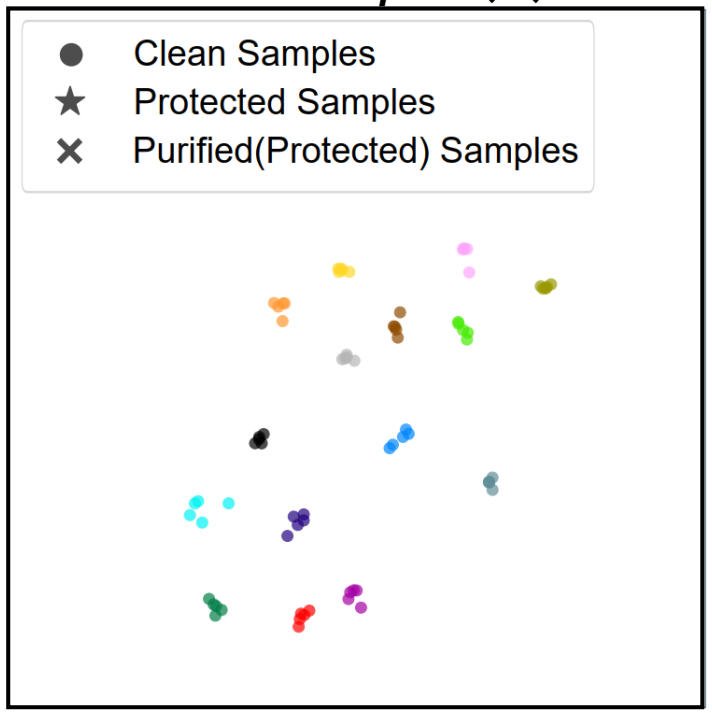
- First systematic evaluation of protective perturbations against voice cloning when attackers try to purify these perturbations.
  - *Reveal that existing defenses may fail.*

# Existing Purification: Effective But Not Good Enough

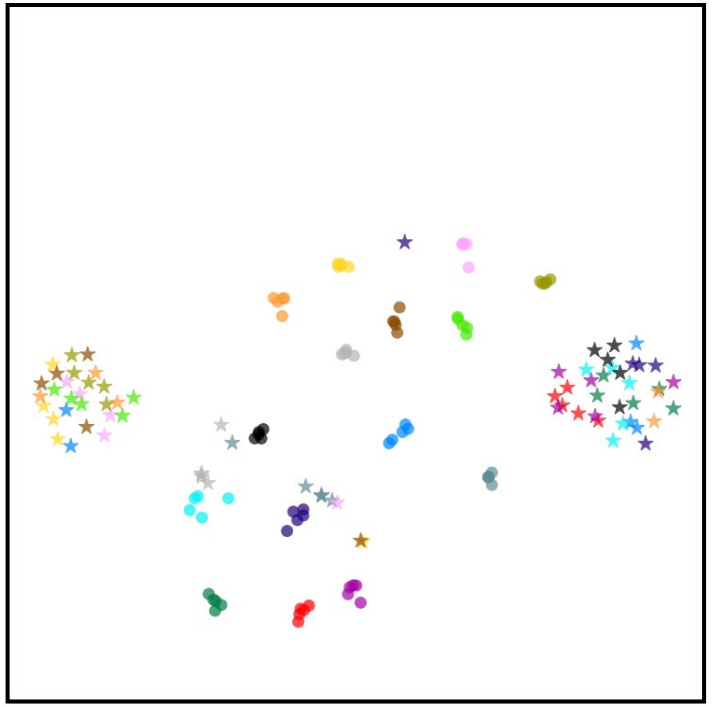


- Most prior purification for *classification tasks*, not voice cloning.
- When applied to voice cloning, they can neutralize *some* protection but...

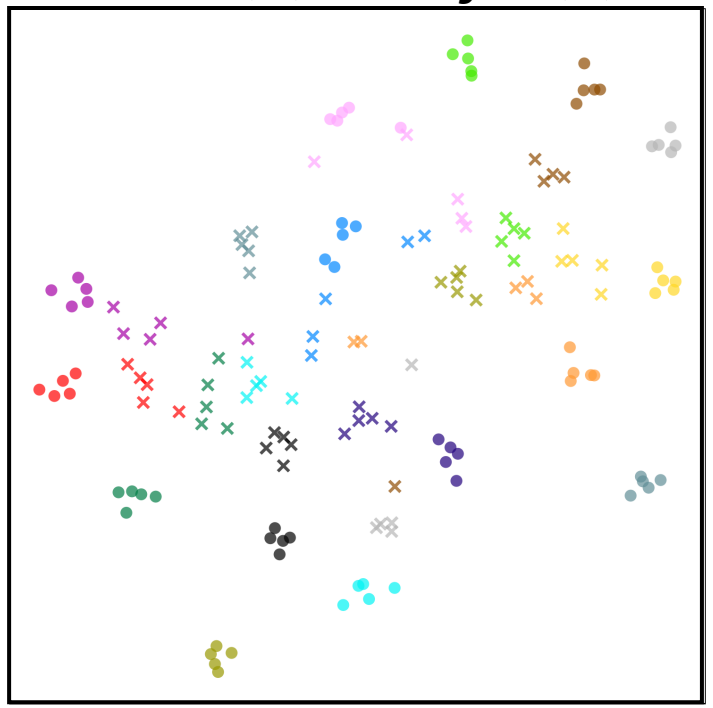
***Clean Samples* (●)**




***Clean* (●) *vs. Protected* (★)**



***Clean* (●) *vs. Purified* (×)**



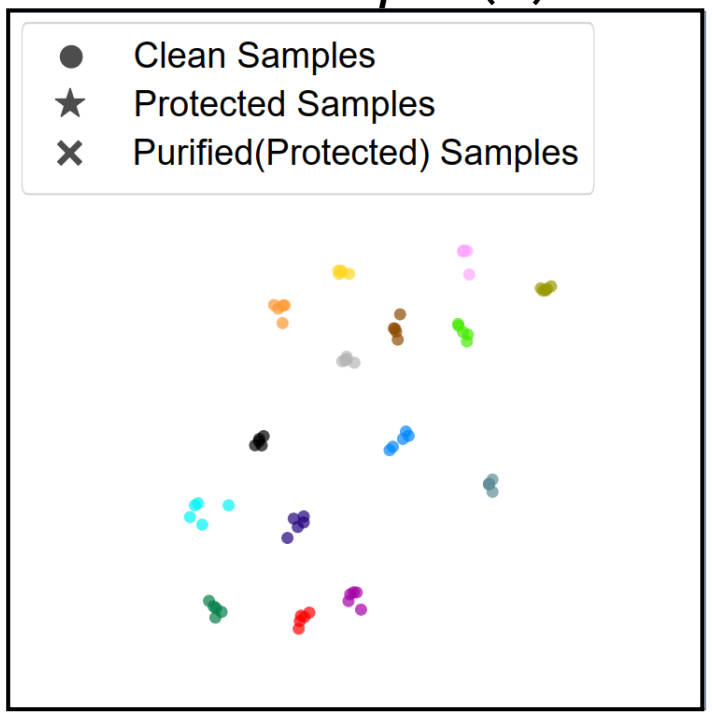
** Different colors (🎨) represents different speakers.*

# Existing Purification: Effective But Not Good Enough

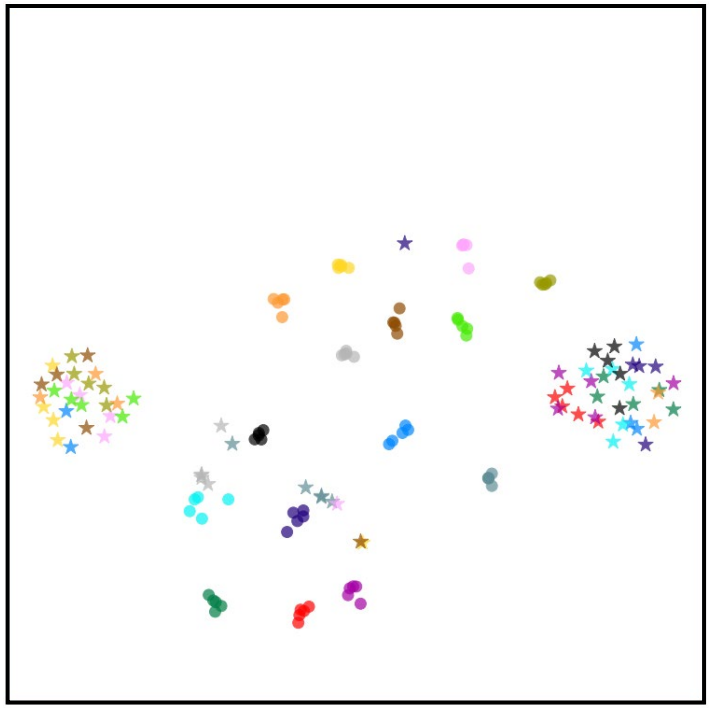


- Most prior purification for *classification tasks*, not voice cloning.
- When applied to voice cloning, they can neutralize *some* protection but...

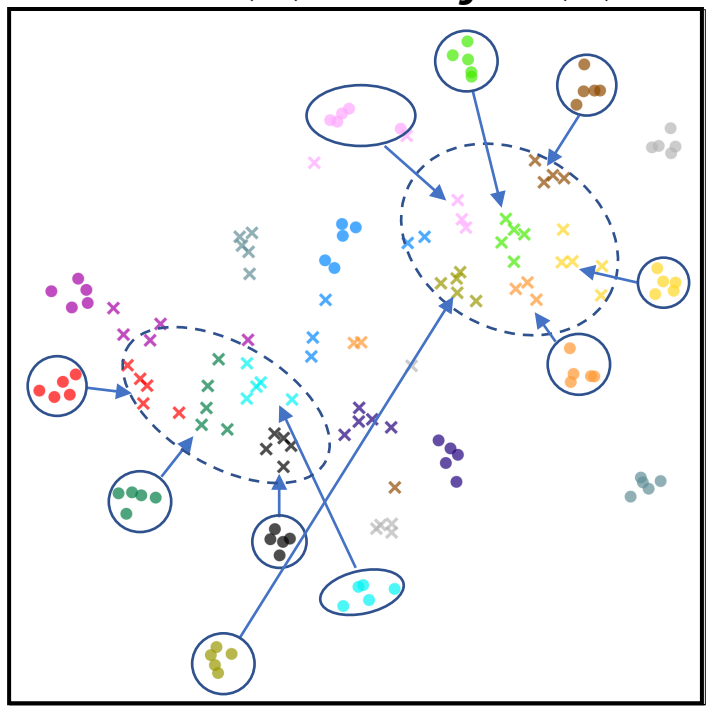
***Clean Samples*** **(●)**




***Clean (●) vs. Protected (★)***



***Clean (●) vs. Purified (×)***



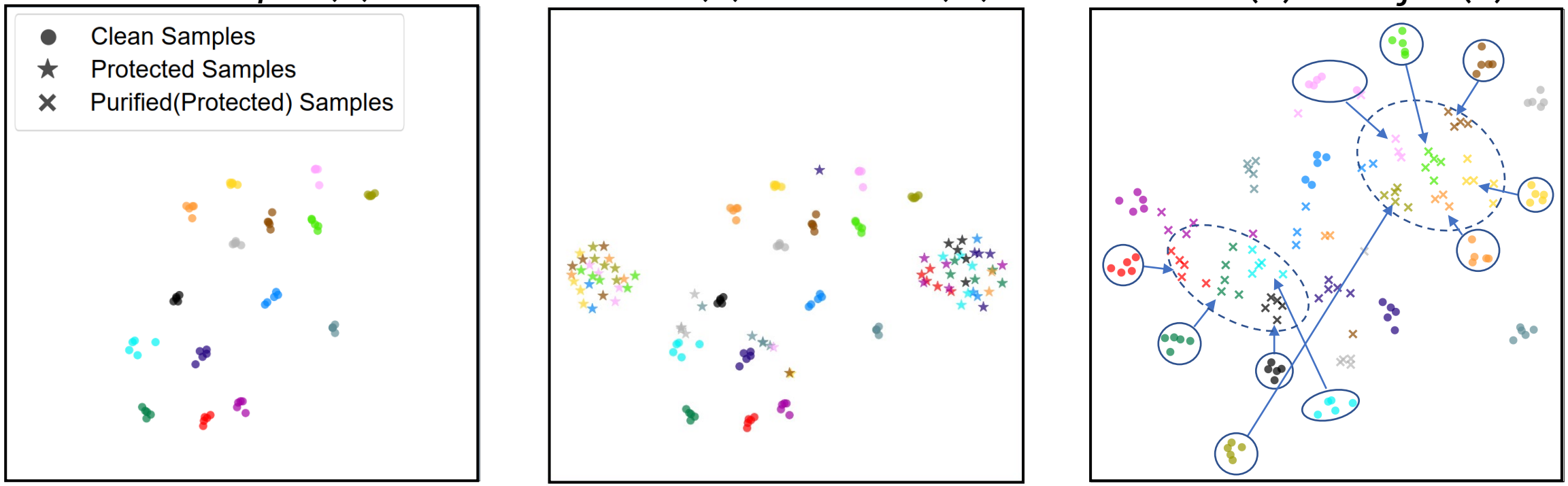
** Different colors (🎨) represents different speakers. Arrows (→) point from clean samples toward their corresponding purified samples.*

# Existing Purification: Effective But Not Good Enough



*Existing purification introduces distortions in voice cloning model embedding spaces, therefore degrade voice cloning performance.*

***Clean Samples (●)*** ***Clean (●) vs. Protected (★)*** ***Clean (●) vs. Purified (×)***




** Different colors (🎨) represents different speakers. Arrows (→) point from clean samples toward their corresponding purified samples.*

# Our Contribution 2



- First systematic evaluation of protective perturbations against voice cloning when attackers try to purify these perturbations.
  - *Reveal that existing defenses may fail.*
- Propose a novel purification method (PhonePuRe) to bypass existing protections.
  - *Outperforms baselines, further exposing risks in existing defenses.*

# Proposed Idea: Purification-Refinement Framework



**PhonePuRe** (**Pu**rification + **Phone**me-Guided **Re**finement)

- **Insight:** Purified distributions deviate from clean ones.
- **Two-Stage Framework:**
  - Purification Stage: Preliminarily mitigate noise (unconditional diffusion).
  - Phoneme-Guided Refinement Stage: Align closer with clean distribution (conditional diffusion).

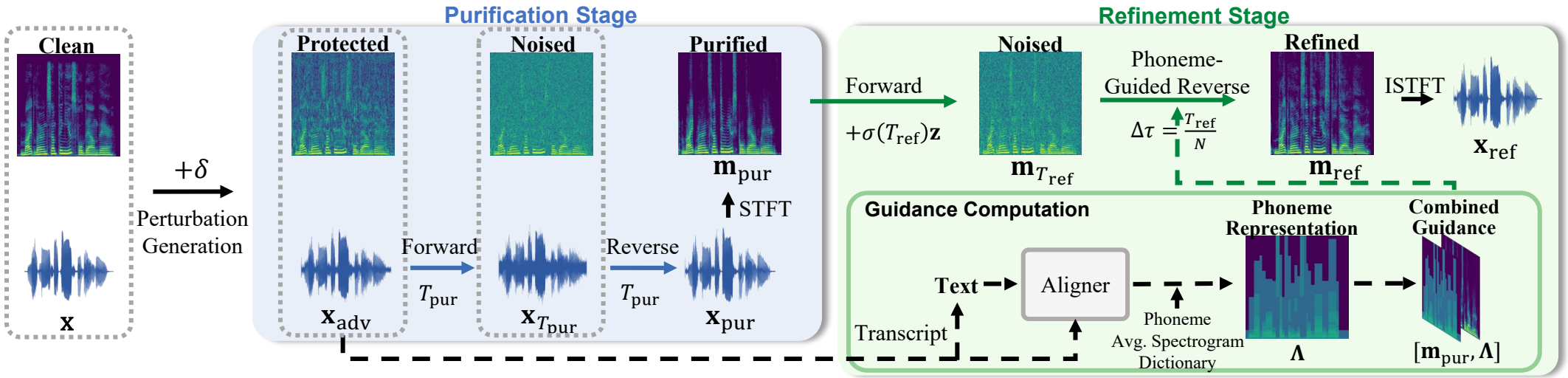



Figure: Inference process of our framework.

# Proposed Idea: Purification-Refinement Framework



**PhonePuRe** (**Pu**rification + **Phone**me-Guided **Re**finement)

- **Insight:** Purified distributions deviate from clean ones.
- **Two-Stage Framework:**
  - Purification Stage: Preliminarily mitigate noise (unconditional diffusion).
  - Phoneme-Guided Refinement Stage: Align closer with clean distribution (conditional diffusion).

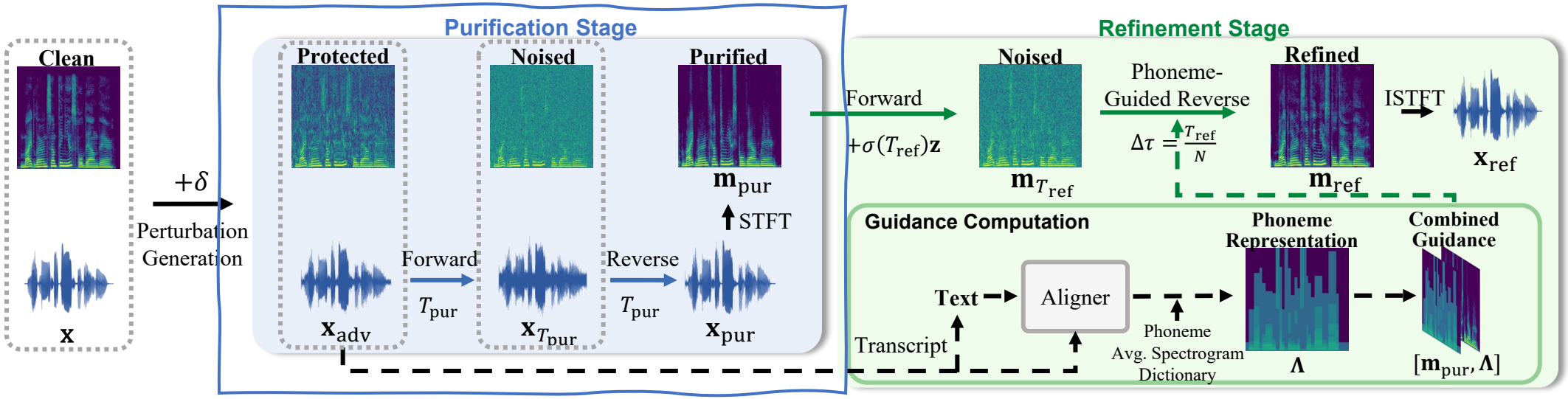



Figure: Inference process of our framework.

# Proposed Idea: Purification-Refinement Framework



**PhonePuRe** (**Pu**rification + **Phone**me-Guided **Re**finement)

- **Insight:** Purified distributions deviate from clean ones.
- **Two-Stage Framework:**
  - Purification Stage: Preliminarily mitigate noise (unconditional diffusion).
  - Phoneme-Guided Refinement Stage: Align closer with clean distribution (conditional diffusion).

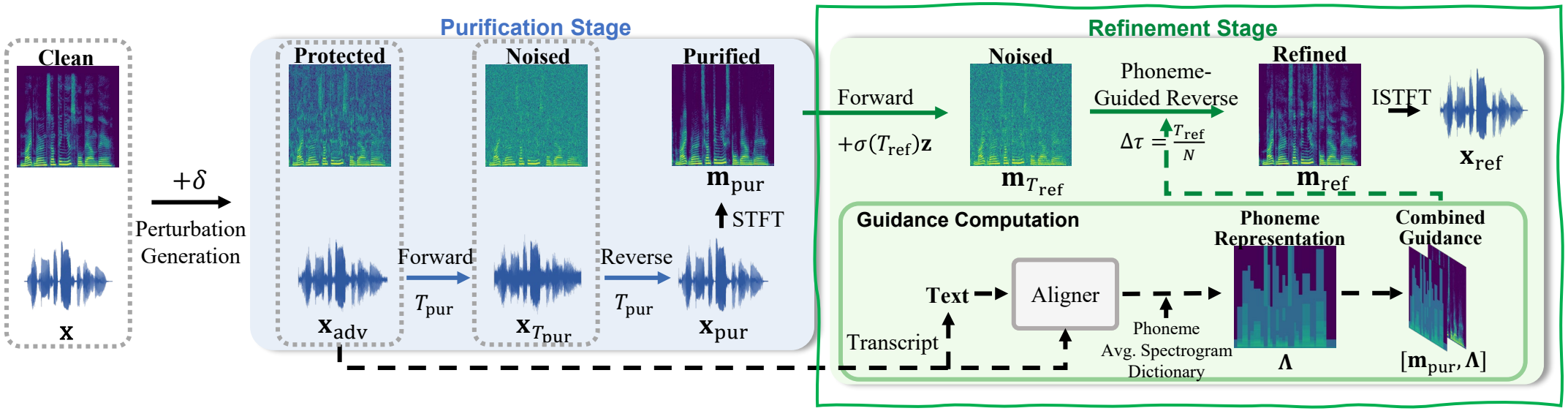
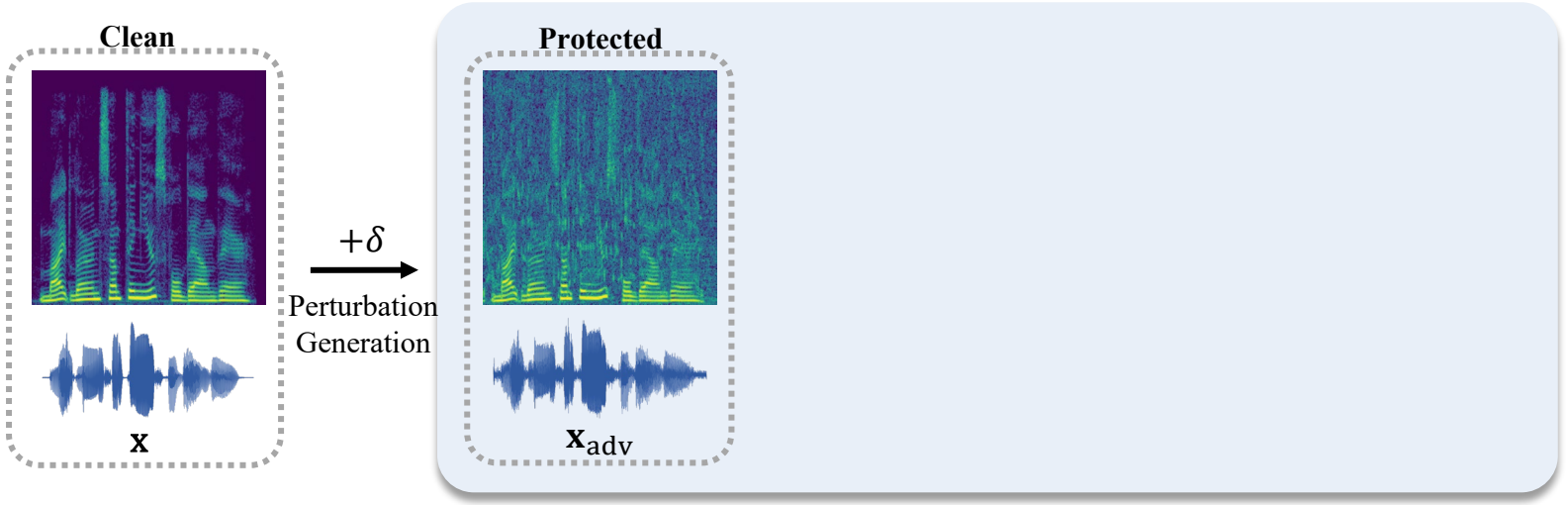



Figure: Inference process of our framework.

# Purification Stage: Unconditional Diffusion



- Employs *DiffWave* model (on waveforms).
- Input: $\mathbf{x}_{\mathrm{adv}}$. Output: $\mathbf{x}_{\mathrm{pur}}$.

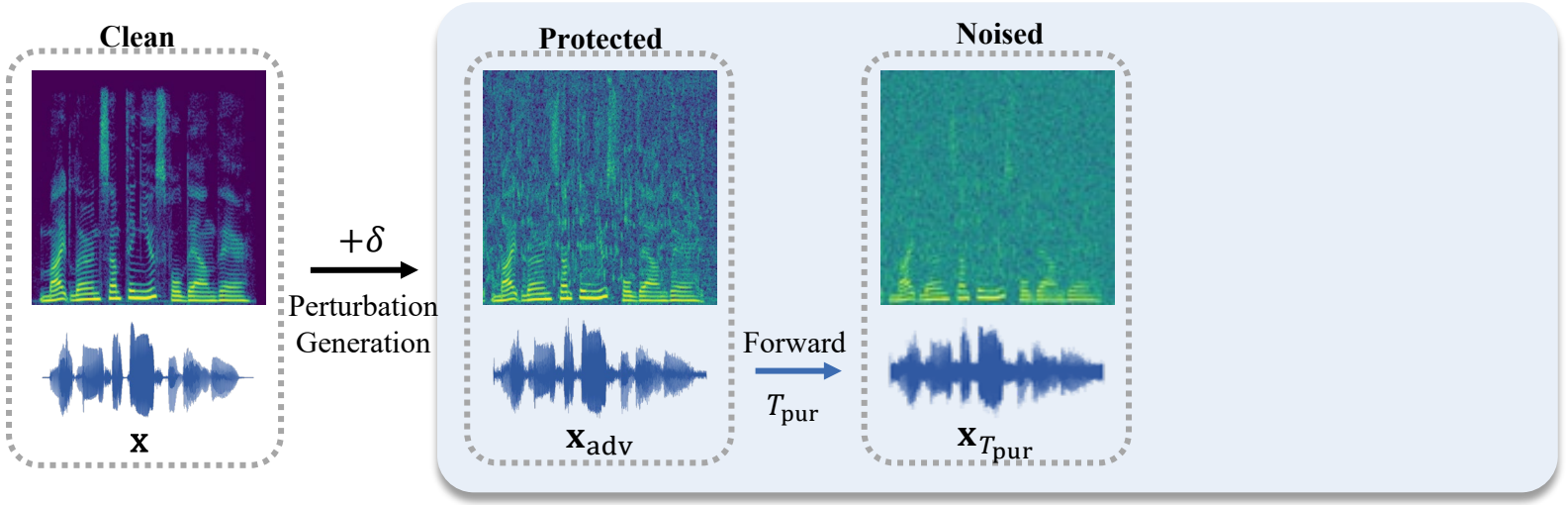


# Purification Stage: Unconditional Diffusion



- Employs *DiffWave* model (on waveforms).
- Input: $\mathbf{x}_{\text{adv}}$. Output: $\mathbf{x}_{\text{pur}}$.
- **Forward Diffusion**: Add noise ($T_{\text{pur}}$ steps).

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right)$$



# Purification Stage: Unconditional Diffusion

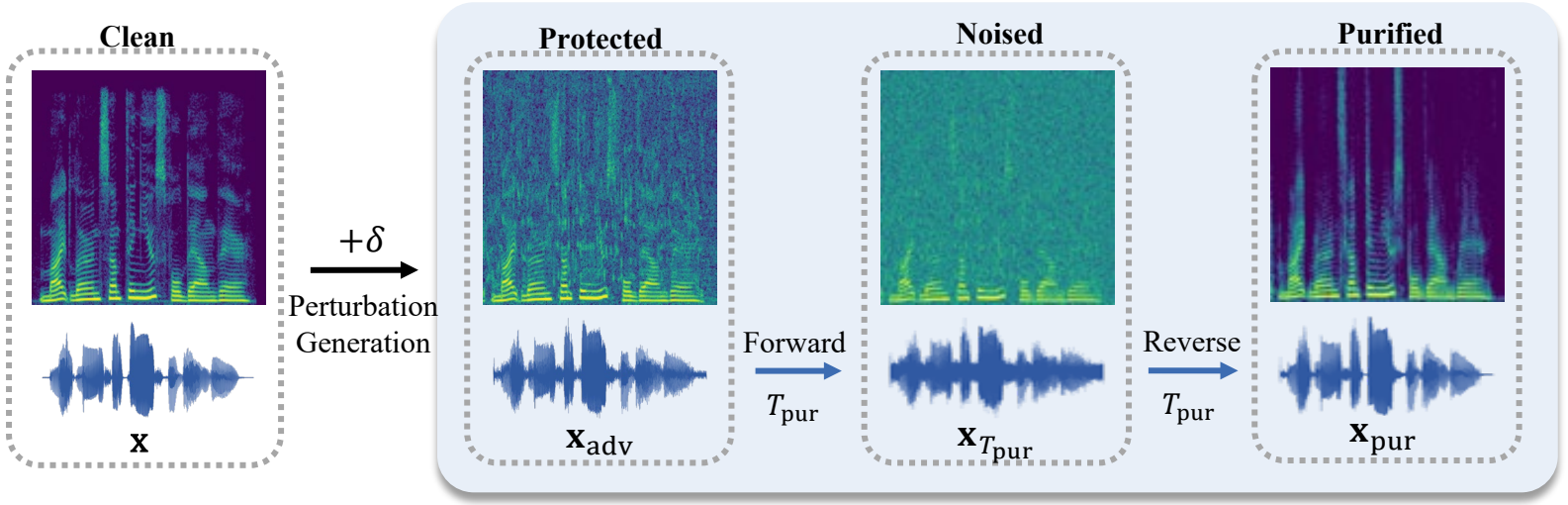


- Employs *DiffWave* model (on waveforms).
- Input: $\mathbf{x}_{\text{adv}}$. Output: $\mathbf{x}_{\text{pur}}$.
- **Forward Diffusion**: Add noise ($T_{\text{pur}}$ steps).

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right)$$

- **Reverse Diffusion**: Denoise ($T_{\text{pur}}$ steps).

$$\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2\mathbf{I})$$

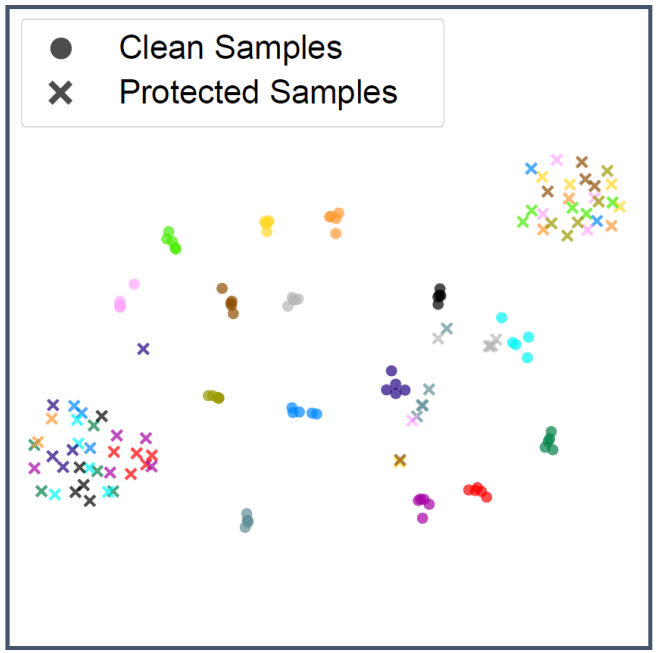


# Refinement Stage: Why Refinement works?



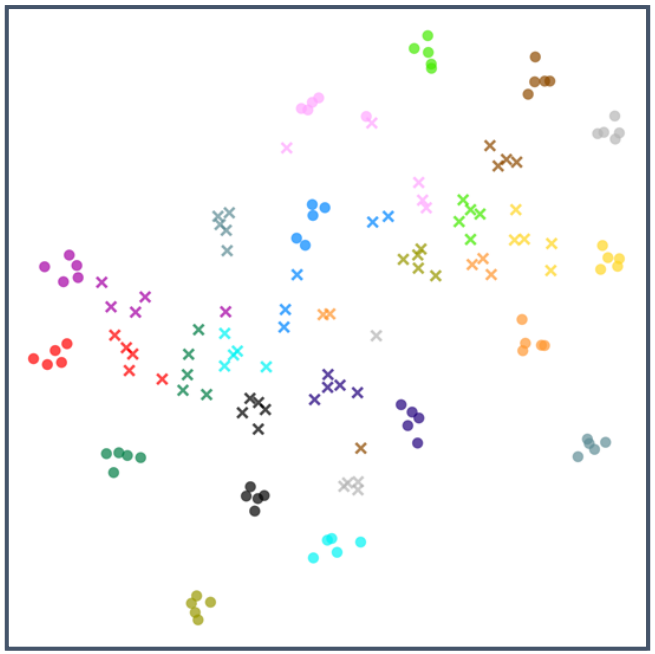
- **Our Observation:** Purified (clean) & Purified (protected) samples have similar distributions.

***Clean* (●) *vs.***
***Protected* (×)**




***Purified***
***Protected***

***Clean* (●) *vs.***
***Purified (Protected)* (×)**

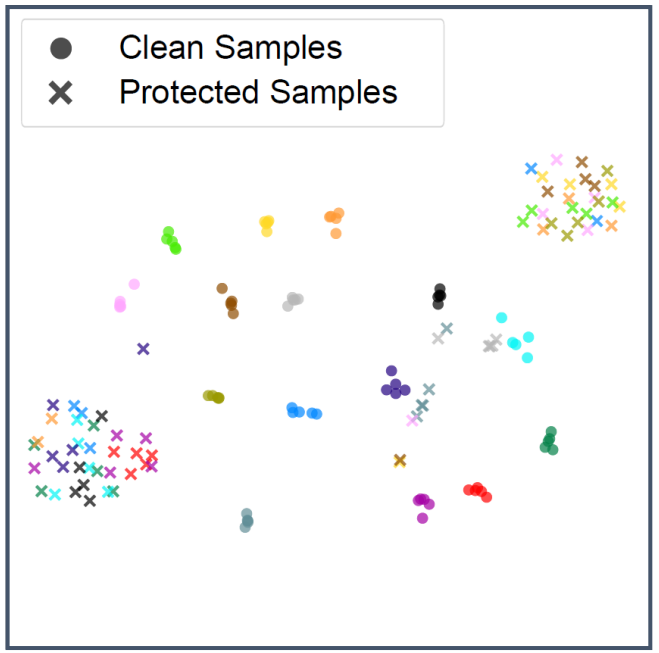


# Refinement Stage: Why Refinement works?



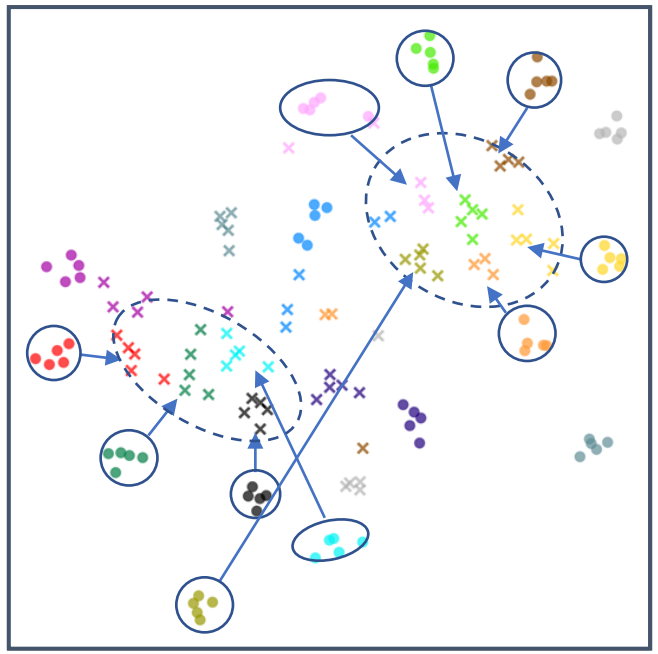
- **Our Observation:** Purified (clean) & Purified (protected) samples have similar distributions.

***Clean* (●) *vs.***
***Protected* (×)**




***Purified***
***Protected***

***Clean* (●) *vs.***
***Purified (Protected)* (×)**

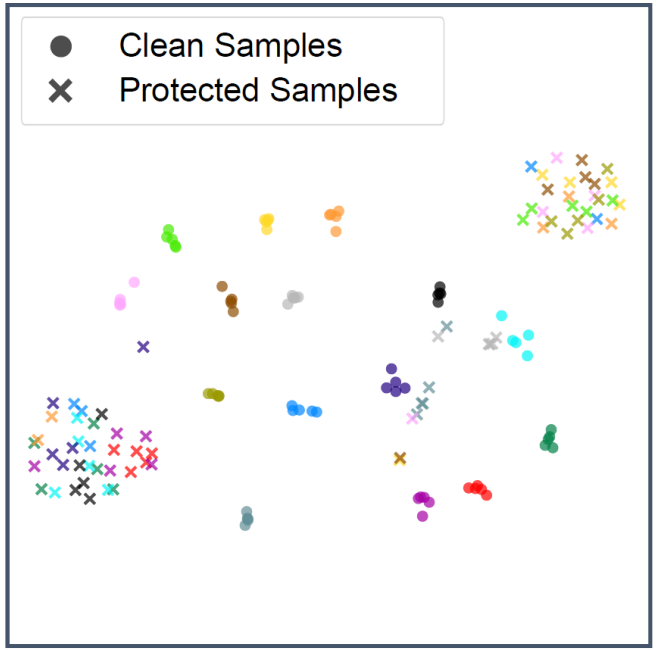


# Refinement Stage: Why Refinement works?



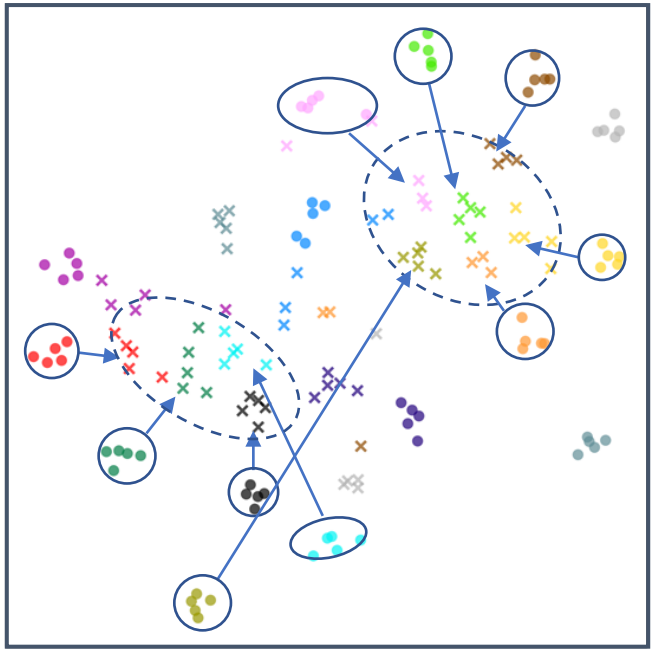
- **Our Observation:** Purified (clean) & Purified (protected) samples have similar distributions.

***Clean* (●) *vs.***
***Protected* (×)**



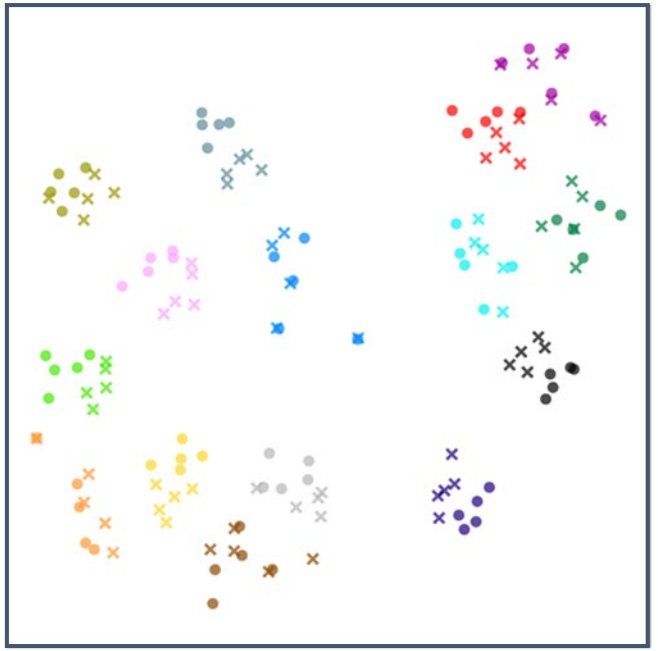

***Purified***
***Protected***

***Clean* (●) *vs.***
***Purified* (*Protected*) (×)**



***Purified***
***Clean***

***Purified* (*Clean*) (●) *vs.***
***Purified* (*Protected*) (×)**

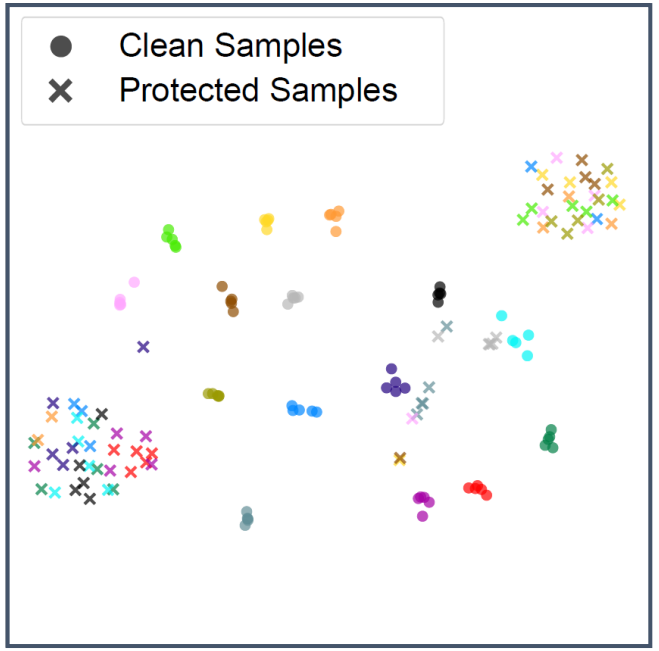


# Refinement Stage: Why Refinement works?



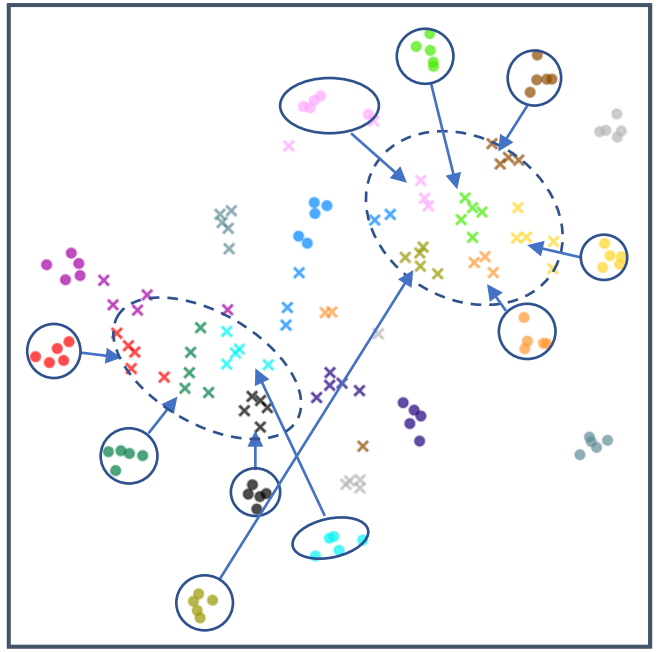
- Therefore, if we train a **Refinement model** to map Purified (clean) to clean, it will be likely to map Purified (protected) to nearly clean distributions.

***Clean* (●) *vs.***
***Protected* (×)**




***Purified***
***Protected***

***Clean* (●) *vs.***
***Purified (Protected)* (×)**

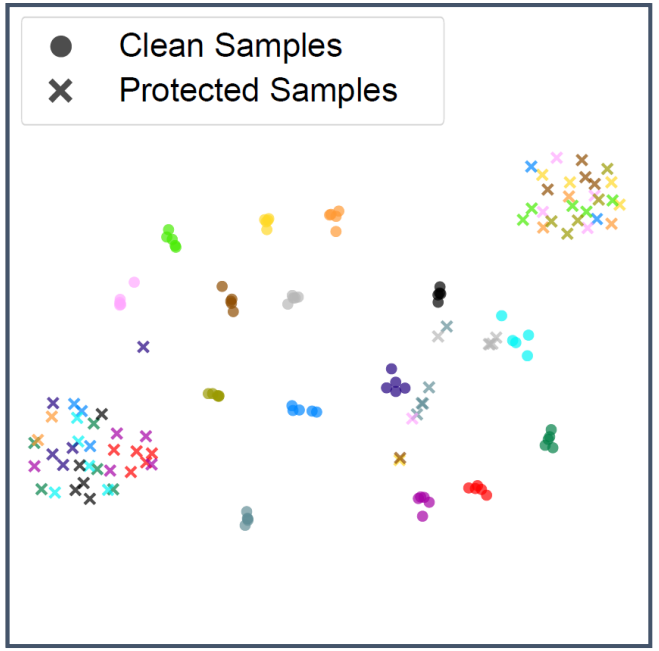


# Refinement Stage: Why Refinement works?



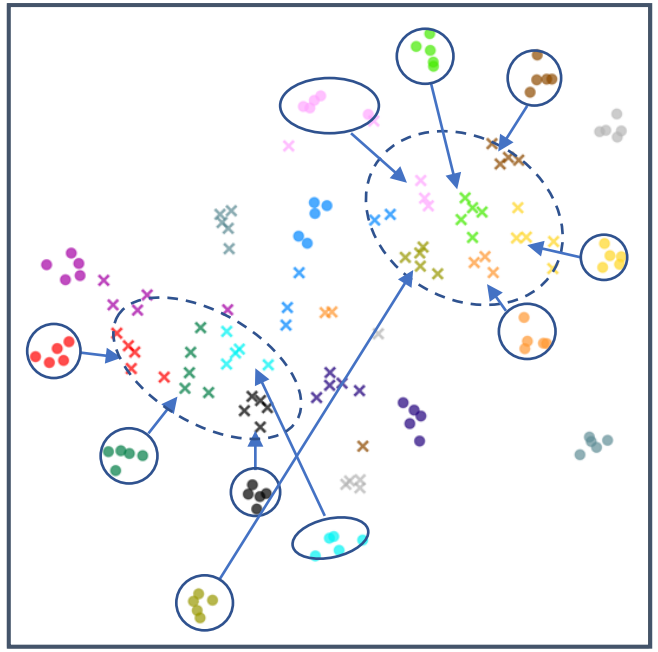
- Therefore, if we train a **Refinement model** to map Purified (clean) to clean, it will be likely to map Purified (protected) to nearly clean distributions.

***Clean (●) vs. Protected (×)***



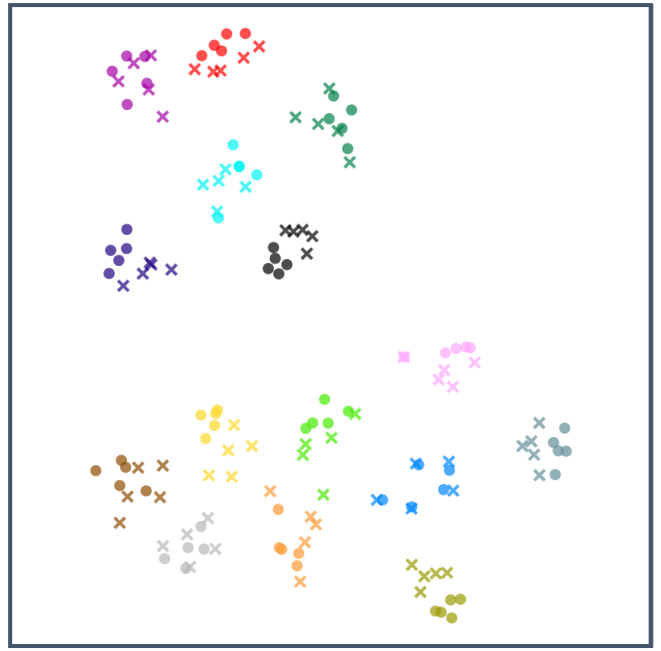

***Purified Protected***

***Clean (●) vs. Purified (Protected) (×)***



***Refined Protected***

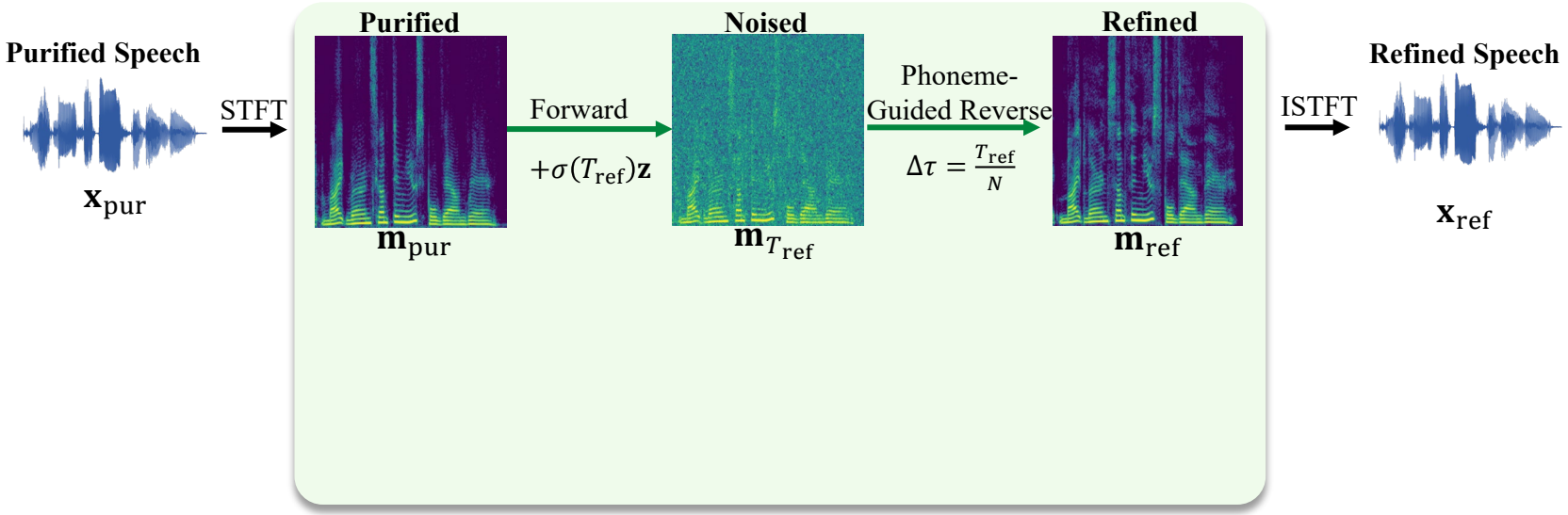
***Clean (●) vs. Refined (Purified (Protected)) (×)***



# Refinement Stage: Phoneme-Guided Score-Based Diffusion



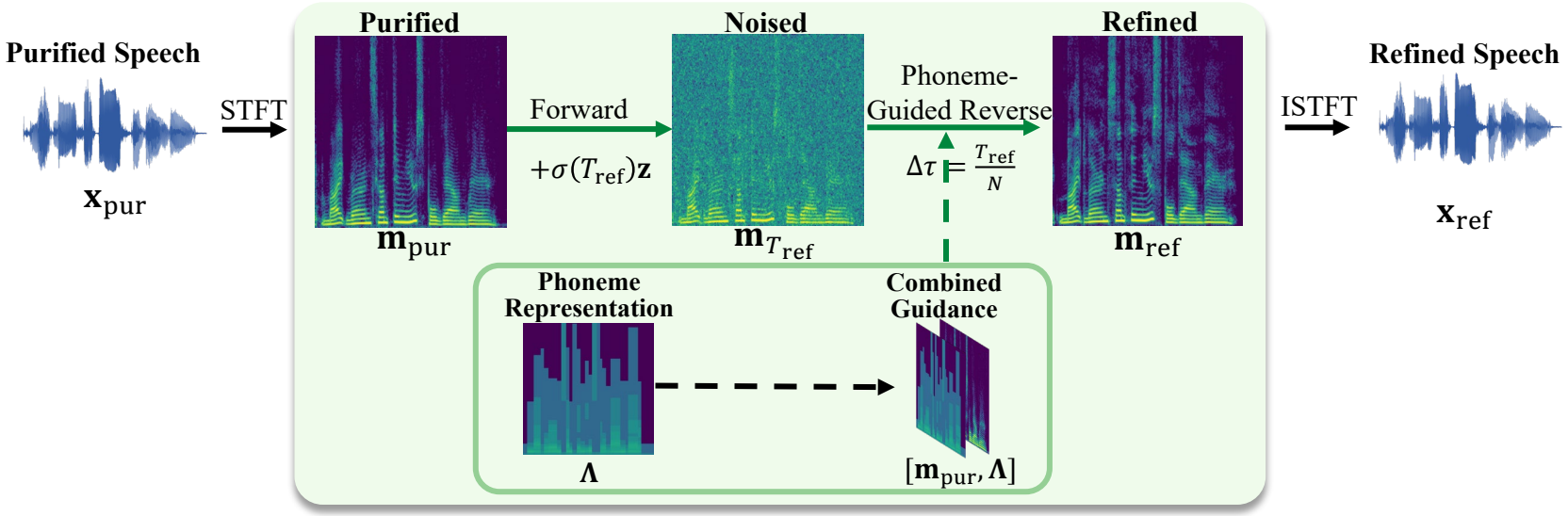
- Employs *score-based diffusion* model on complex spectrums ($\mathbf{m} = \mathrm{STFT}(\mathbf{x})$).
- Input: $\mathbf{x}_{\mathrm{pur}}$. Output: $\mathbf{x}_{\mathrm{ref}}$.



# Refinement Stage: Phoneme-Guided Score-Based Diffusion



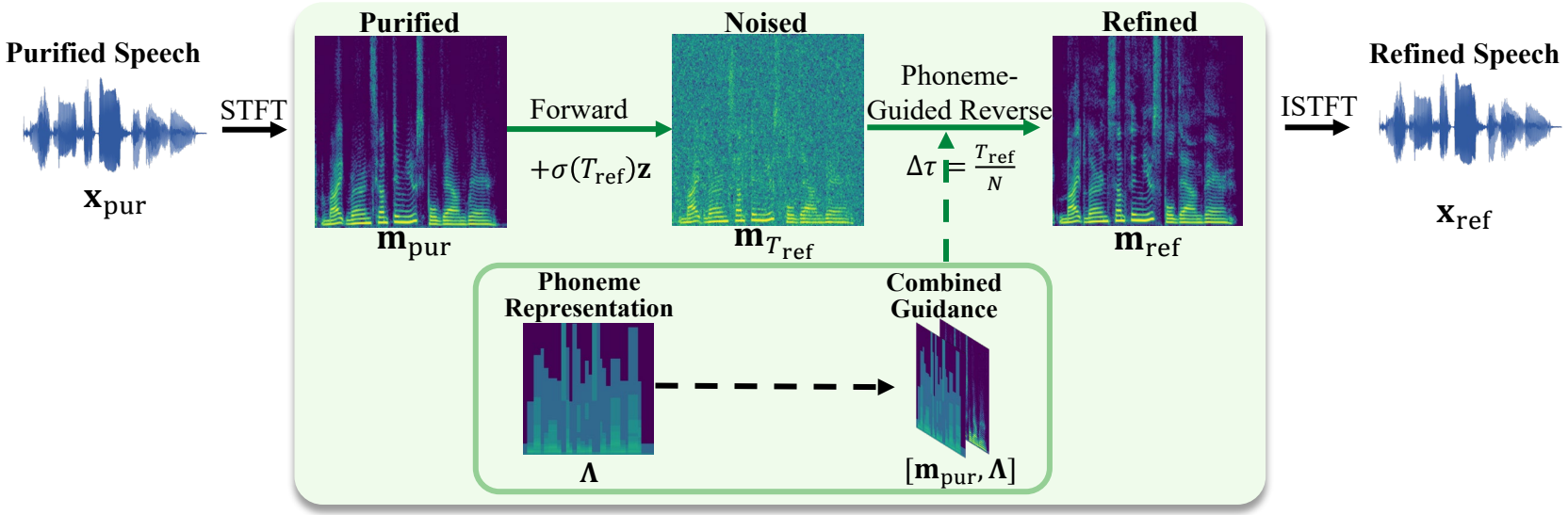
- Employs *score-based diffusion* model on complex spectrums ($\mathbf{m} = \mathrm{STFT}(\mathbf{x})$).
- Input: $\mathbf{x}_{\mathrm{pur}}$. Output: $\mathbf{x}_{\mathrm{ref}}$.
- Use Phoneme Guidance $\mathbf{\Lambda}$.



# Refinement Stage: Phoneme-Guided Score-Based Diffusion



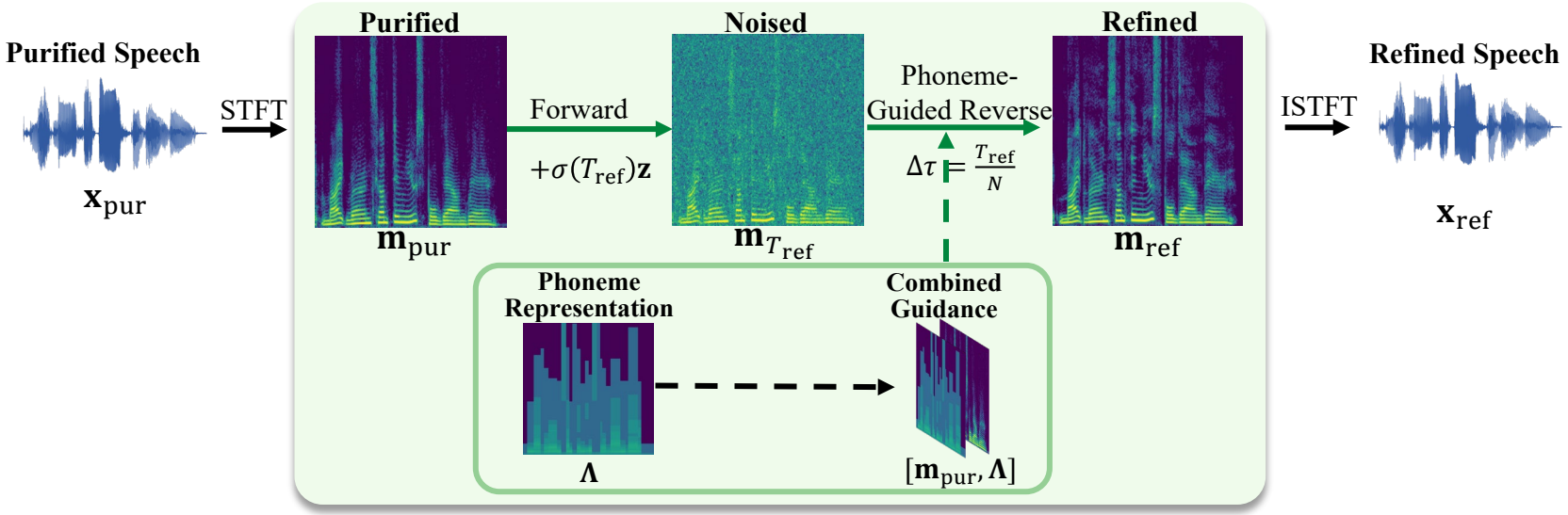
- Employs *score-based diffusion* model on complex spectrums ($\mathbf{m} = \mathrm{STFT}(\mathbf{x})$).
- Input: $\mathbf{x}_{\mathrm{pur}}$. Output: $\mathbf{x}_{\mathrm{ref}}$.
- Use Phoneme Guidance $\mathbf{\Lambda}$.
- **Training:** Learns to generate $\mathbf{m}_{\mathrm{clean}}$ from $\mathbf{m}_{\mathrm{pur}}$ (Purified (clean) samples).



# Refinement Stage: Phoneme-Guided Score-Based Diffusion



- Employs *score-based diffusion* model on complex spectrums ($\mathbf{m} = \mathrm{STFT}(\mathbf{x})$).
- Input: $\mathbf{x}_{\mathrm{pur}}$. Output: $\mathbf{x}_{\mathrm{ref}}$.
- Use Phoneme Guidance $\mathbf{\Lambda}$.
- **Training:** Learns to generate $\mathbf{m}_{\mathrm{clean}}$ from $\mathbf{m}_{\mathrm{pur}}$ (Purified (clean) samples).
- **Inference:** Generates $\mathbf{m}_{\mathrm{ref}}$ from $\mathbf{m}_{\mathrm{pur}}$ (Purified (protected) samples).



# Experimental Setup

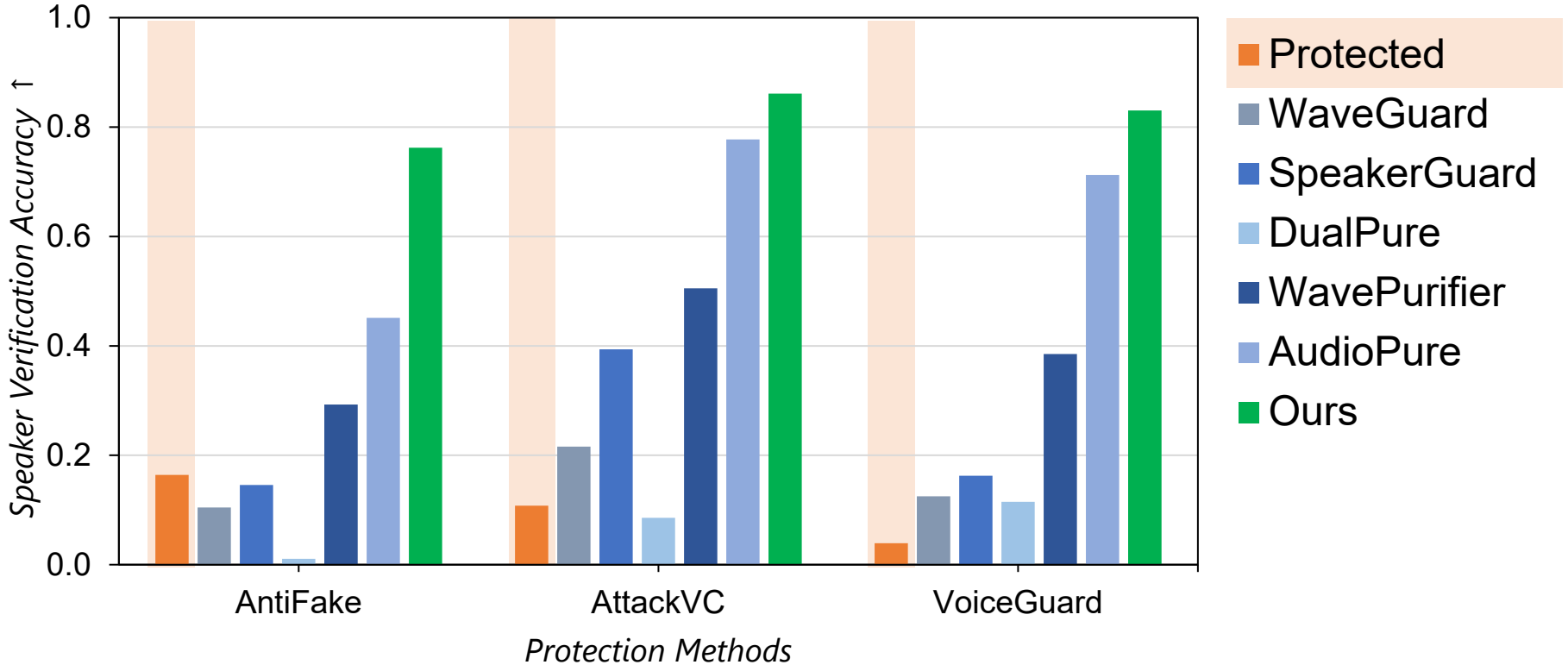


- **Voice Cloning Methods** (6 total)
  - TTS: YourTTS, SV2TTS, Tortoise
  - Voice Conversion: DiffVC, OpenVoice V2, SeedVC
- **Protection Methods Evaluated** (3 total)
  - AntiFake, AttackVC, VoiceGuard
- **Adversarial Purification Baselines** (5 total)
  - Transformation-based: WaveGuard, SpeakerGuard
  - Reconstruction-based: AudioPure, WavePurifier, DualPure

# Experiment: Objective Results on Effectiveness



- **Existing protection:** Effective w/o purification

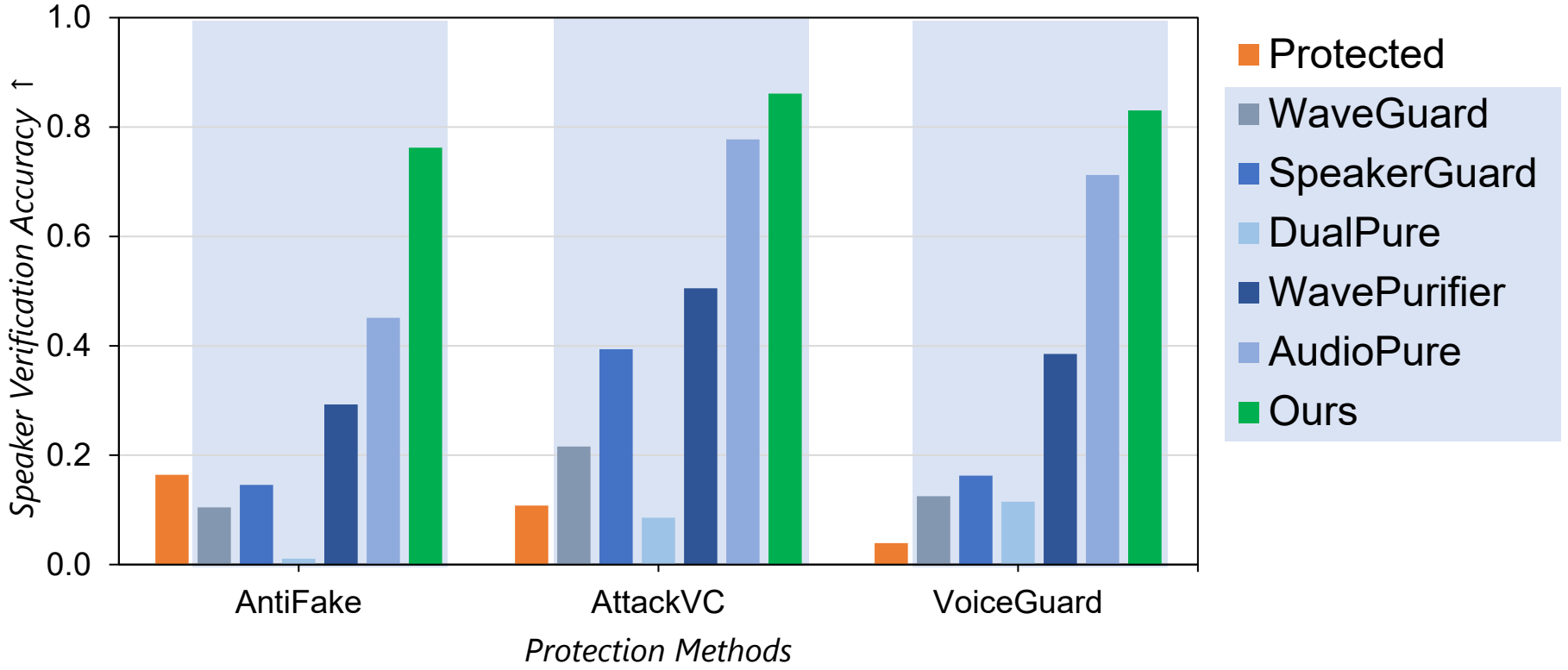



- *Lower SVA mains effective protection.*

# Experiment: Objective Results on Effectiveness



- **Existing protection:** Effective w/o purification; but vulnerable to purification.

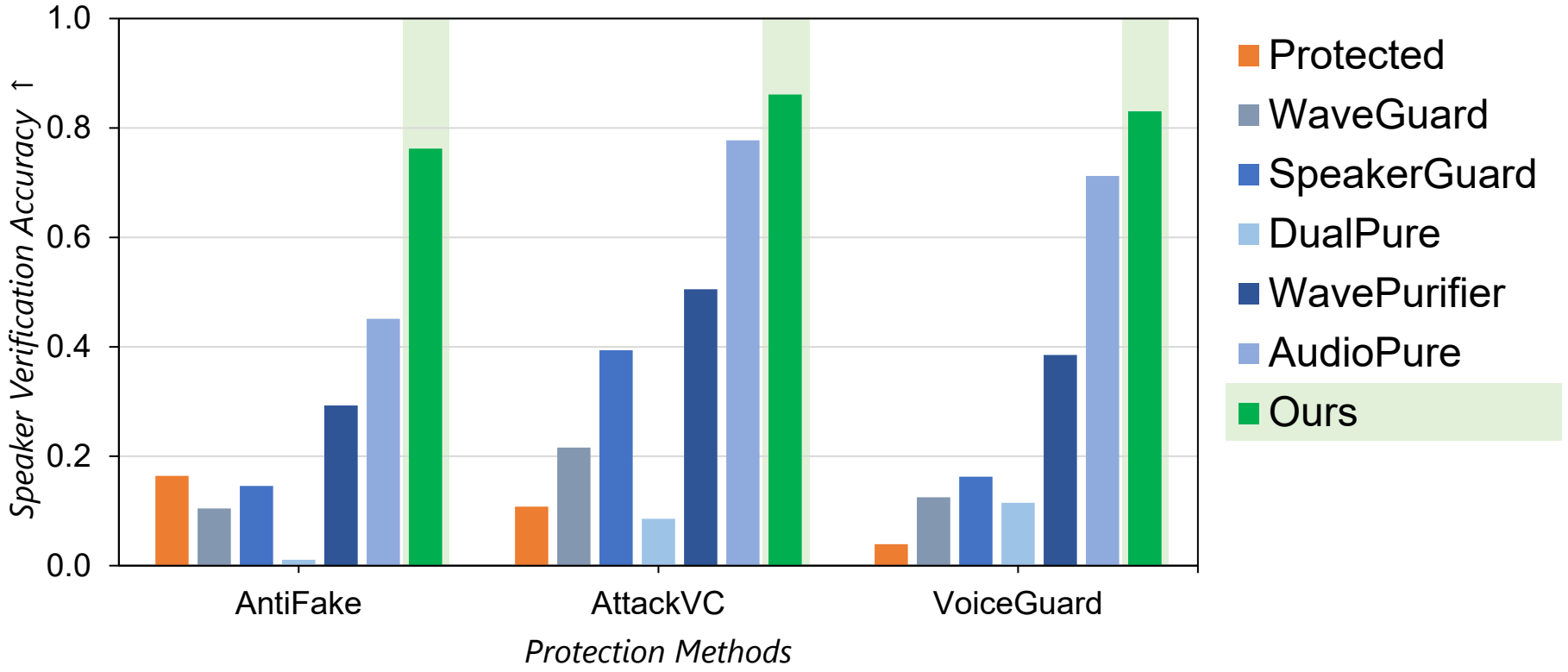



- *Lower SVA mains effective protection.*

# Experiment: Objective Results on Effectiveness



- **Existing protection:** Effective w/o purification; but vulnerable to purification.

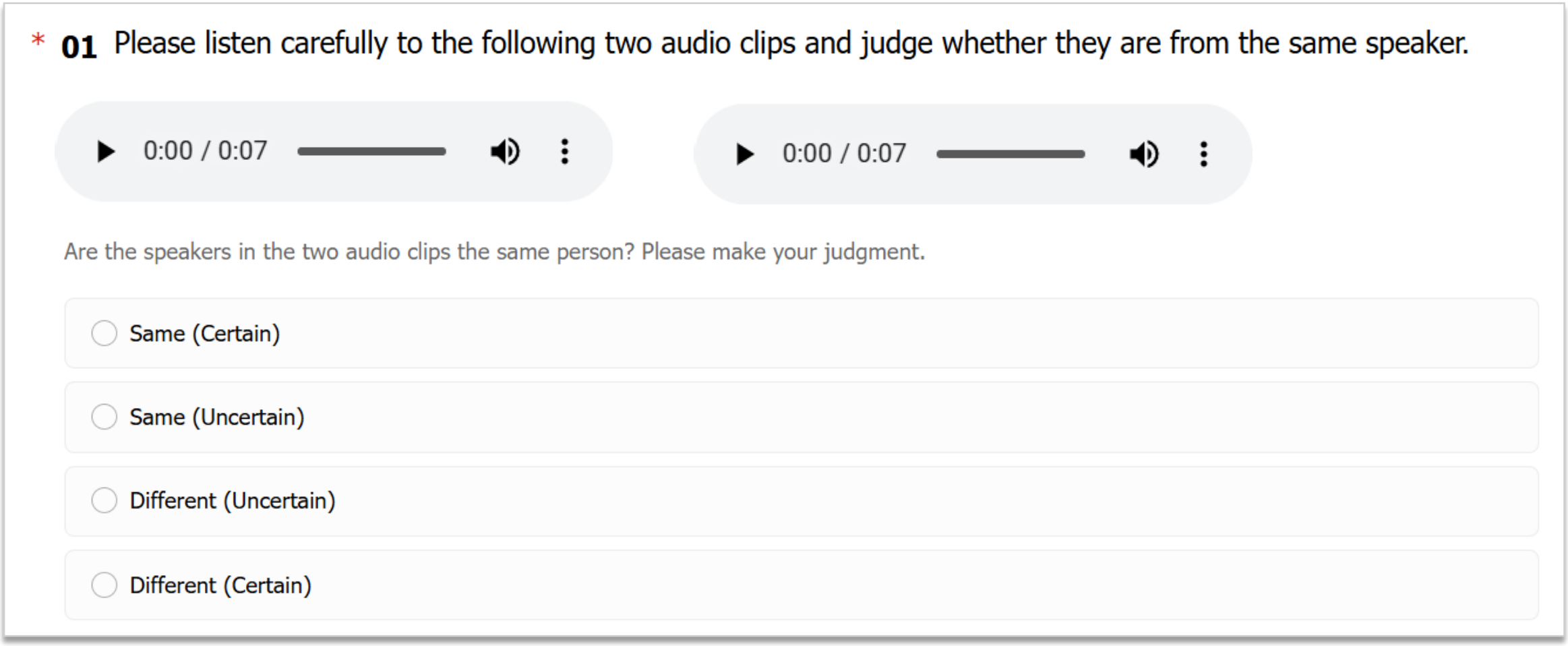



☑ *Our PhonePuRe purification outperforms baselines in bypassing SV.*

# Experiment: Subjective Results on Effectiveness



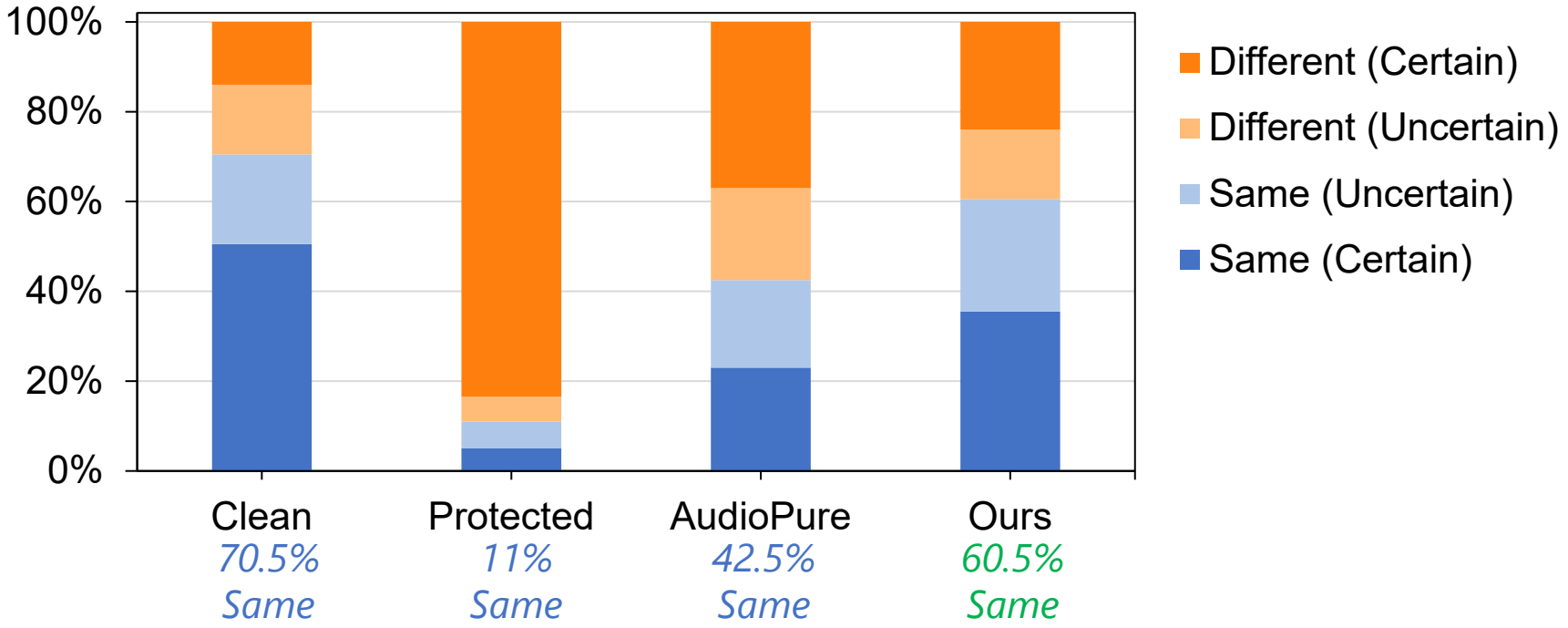
- **Subjective Metric:** Human listening test (*perceived speaker similarity*)



# Experiment: Subjective Results on Effectiveness



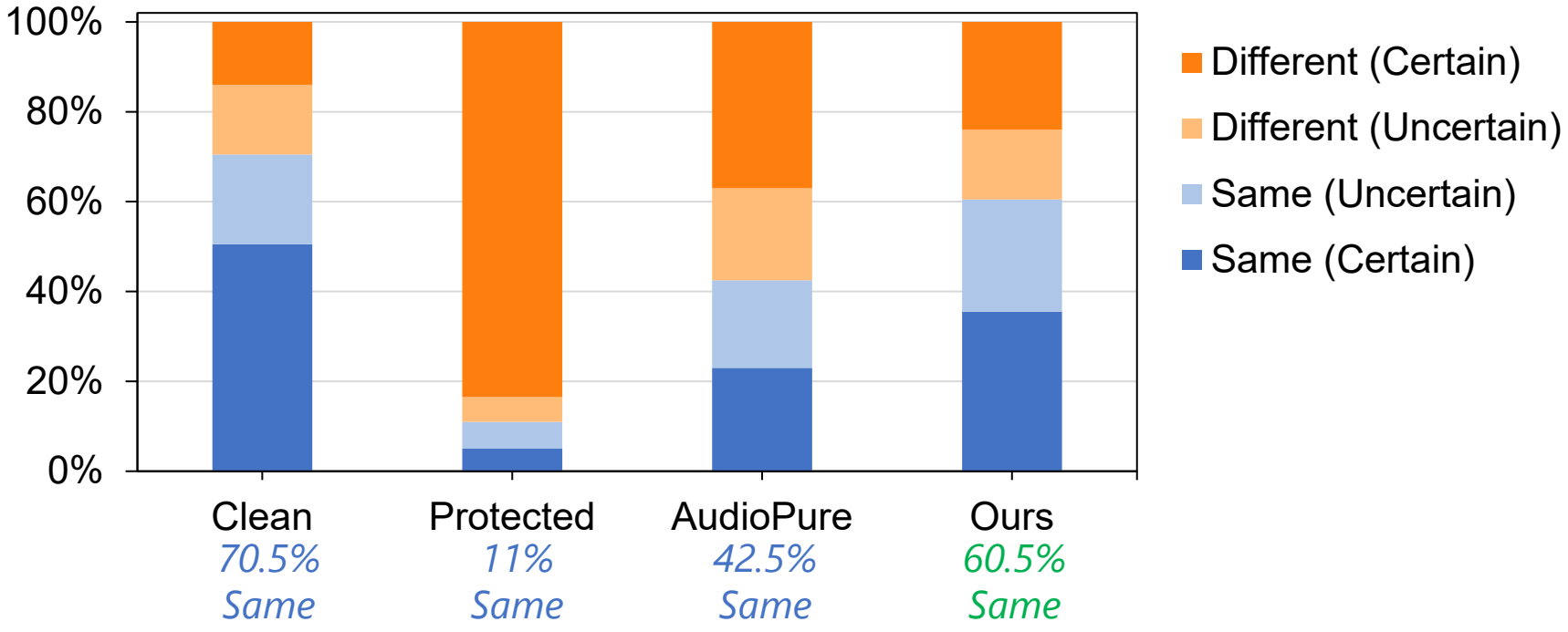
- **Existing protection:** Effective w/o purification (11% Same); but vulnerable to purification (42.5% Same).



# Experiment: Subjective Results on Effectiveness



- **Existing protection:** Effective w/o purification (11% Same); but vulnerable to purification (42.5% Same).




☑ *PhonePuRe bypasses SV & show higher human-perceived similarity (60.5% Same).*

# Our Contribution 3

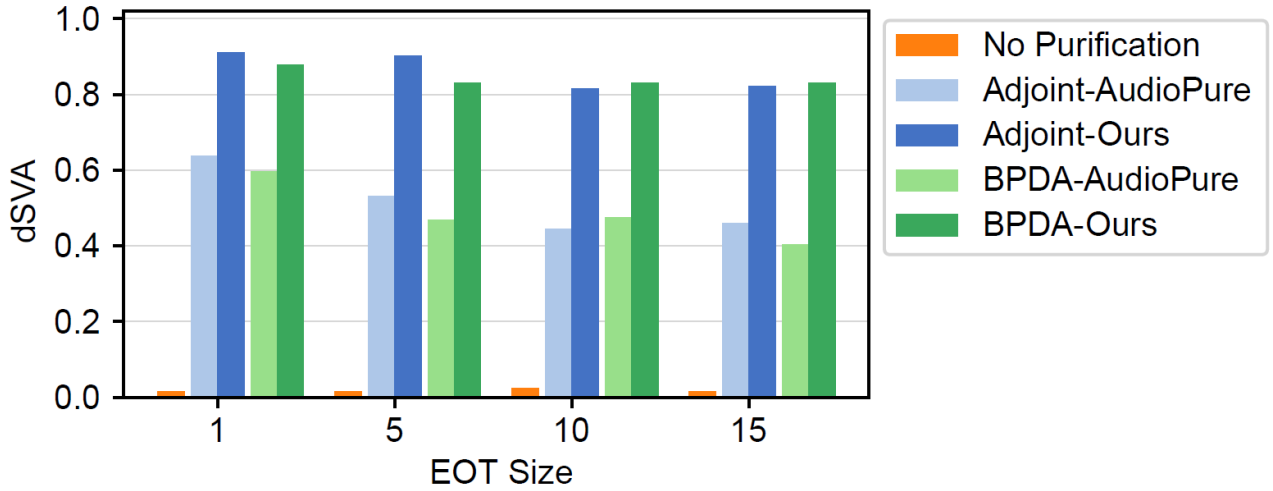


- First systematic evaluation of protective perturbations against voice cloning when attackers try to purify these perturbations.
  - *Reveal that existing defenses may fail.*
- Propose a novel purification method (PhonePuRe) to bypass existing protections.
  - *Outperforms baselines, further exposing risks in existing defenses.*
- Evaluate robustness of our purification against adaptive protections.
  - *Show generating effective defenses against our method is challenging.*

# Experiment: Robustness Against Adaptive Protection



- **Adaptive Protection:** Protector designs perturbations *considering protection*.
- **Challenge:** Calculating the gradients of diffusion models is hard.
- **Two gradient approximation strategies:** BPDA (+EOT), Adjoint (+EOT)

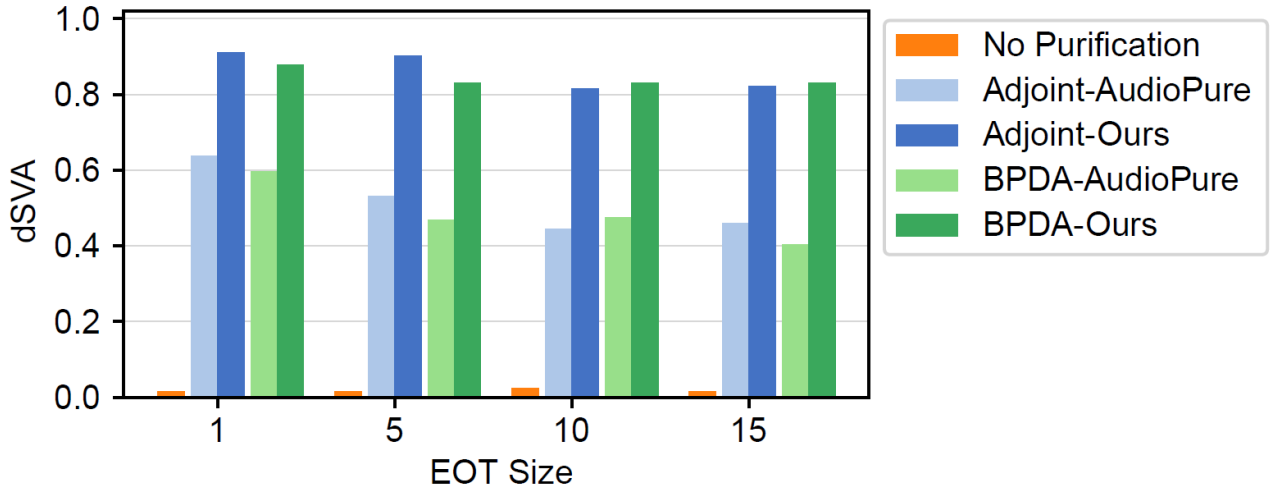



- 80% cloned samples synthesized from our purified samples successfully bypass SV.

# Experiment: Robustness Against Adaptive Protection



- **Adaptive Protection:** Protector designs perturbations *considering protection*.
- **Challenge:** Calculating the gradients of diffusion models is hard.
- **Two gradient approximation strategies:** BPDA (+EOT), Adjoint (+EOT)




- 80% cloned samples synthesized from our purified samples successfully bypass SV.

☑ *Protectors struggle to generate effective perturbations even in white-box scenarios.*

# Conclusion



- First systematic evaluation of protective perturbations against voice cloning when attackers try to purify these perturbations.
  - *Reveal that existing defenses may fail.*
- Propose a novel purification method (PhonePuRe) to bypass existing protections.
  - *Outperforms baselines, further exposing risks in existing defenses.*
- Evaluate robustness of our purification against adaptive protections.
  - *Show generating effective defenses against our method is challenging.*

# Conclusion



- First systematic evaluation of protective perturbations against voice cloning when attackers try to purify these perturbations.
  - *Reveal that existing defenses may fail.*
- Pro
  prot
  -
- Evaluate robustness of our purification against adaptive protections.
  - *Show generating effective defenses against our method is challenging.*

**Underscore the urgent need for more robust solutions to protect our voice**

中国科学技术大学
University of Science and Technology of China

ICML
International Conference
On Machine Learning

# THANK YOU!

**Demo and code website:** https://de-antifake.github.io

**Contact with any questions:** range@mail.ustc.edu.cn