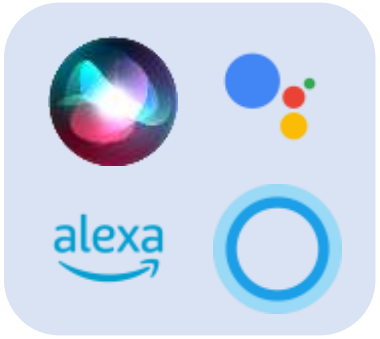
SpeechGuard: Recoverable and Customizable Speech Privacy Protection

Jingmiao Zhang, Suyuan Liu, Jiahui Hou, Zhiqiang Wang,
Haikuo Yu, Xiang-Yang Li

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

Background

- Speech data is everywhere.



Voice Assistants

Online Meetings

Social Media

Smart Cars

Background

- Privacy in Speech 

  - Acoustic Privacy (voiceprints that identify the speaker) ;
  - Content Privacy (sensitive sematic information such as names, addresses, and phone numbers).

Background

- Privacy in Speech 

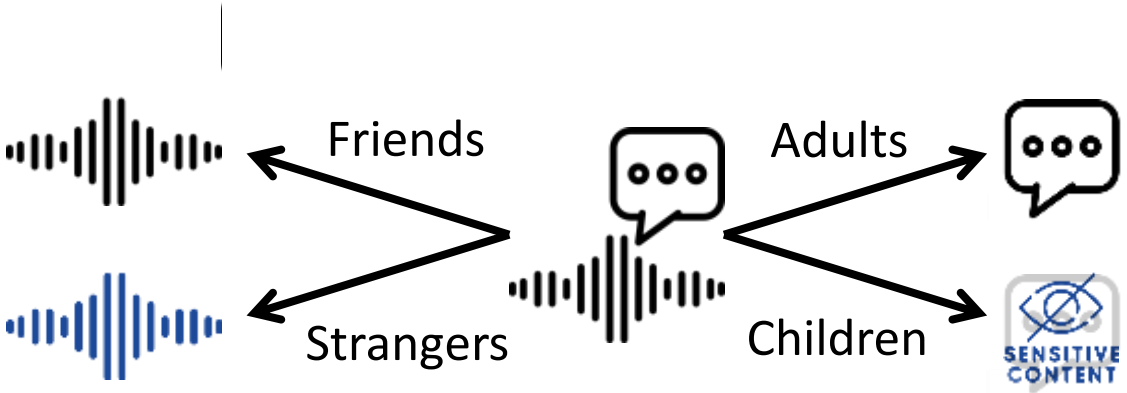
  - Acoustic Privacy (voiceprints that identify the speaker) ;
  - Content Privacy (sensitive sematic information such as names, addresses, and phone numbers).
- Why is Protection Needed?
  - Sharing or processing speech in the cloud exposes it to risks of eavesdropping, misuse, and leakage to third parties.
  - Leaked data can enable voice cloning and sensitive information theft, which may lead to fraud.

Background

- Why is Recoverability Needed?
  - Audio owners want to preserve all information in the cloud without permanent loss.
  - Law enforcement and courts require untampered recordings as authentic evidence.

Background

- Why is Customizability Needed?
  - A single audio file often contains both sensitive and non-sensitive information.
  - Users want the cloud to process the non-sensitive part while protecting the sensitive part.
  - Different listeners need different access: e.g., share the real voice with friends, a protected version with strangers, or restrict content for children while giving adults full access.



Existing Methods

- Irrecoverable: Once sensitive parts are replaced or removed, the original information is permanently lost.
- Non-customizable: Apply the same protection to all listeners without differentiation.

Method	Privacy Protection		Recoverablity		Customizability
	Acoustic	Content	Acoustic	Content	
McAdams (Interspeech'20)	✓	×	✓	×	×
Preech (USENIX Security'20)	✓	✓	×	×	×
VoiceMask (TDSC'21)	✓	✓	✓	×	×
Overo (CCS'22)	✓	✓	✓	×	×
SpeechGuard (Ours)	✓	✓	✓	✓	✓

N. Tomashenko et al. "Introducing the Voice Privacy Initiative." Interspeech 2020.
S. Ahmed et al. "Preech: A system for Privacy-Preserving speech transcription." USENIX Security 2020.
J. Qian et al. "Speech sanitizer: Speech content desensitization and voice anonymization." TDSC 2021.
J. Lim et al. "Overo: Sharing private audio recordings." CCS 2022.

Goal

- Protect privacy while maintaining speech quality, and allow speech to be fully or partially recovered under permissions.

Goal

- Protect privacy while maintaining speech quality, and allow speech to be fully or partially recovered under permissions.
  - Privacy–Quality Trade-off
    - Reduce Automatic Speaker Verification (ASV) and sensitive Automatic Speech Recognition (ASR) accuracy in protected speech.
    - Preserve ASR accuracy for non-sensitive content.

Goal

- Protect privacy while maintaining speech quality, and allow speech to be fully or partially recovered under permissions.
  - Privacy–Quality Trade-off
    - Reduce Automatic Speaker Verification (ASV) and sensitive Automatic Speech Recognition (ASR) accuracy in protected speech.
    - Preserve ASR accuracy for non-sensitive content.
  - Recoverability
    - Within permissions, recover acoustic privacy and content privacy close to the original.

Goal

- Protect privacy while maintaining speech quality, and allow speech to be fully or partially recovered under permissions.
  - Privacy–Quality Trade-off
    - Reduce Automatic Speaker Verification (ASV) and sensitive Automatic Speech Recognition (ASR) accuracy in protected speech.
    - Preserve ASR accuracy for non-sensitive content.
  - Recoverability
    - Within permissions, recover acoustic privacy and content privacy close to the original.
  - Customizability
    - Owner defines which parts to protect and the protection strength.
    - Listeners recover different information based on assigned permissions.

Threat Model

- Roles and Permissions
  - Audio Owner publishes privacy-processed speech.
  - L_1 listeners: can recover all private information (equivalent to owner).
  - L_2 listeners: can recover partial private information, depending on permissions.
  - L_3 listeners: cannot recover any private information.

Threat Model

- Adversaries and Assumptions
  - Adversary capabilities:
    - Access to published protected audio.
    - Guess protection parameters and keys.
    - Use Automatic ASV to infer speaker identity.
    - Use Automatic ASR to infer protected content.

Threat Model

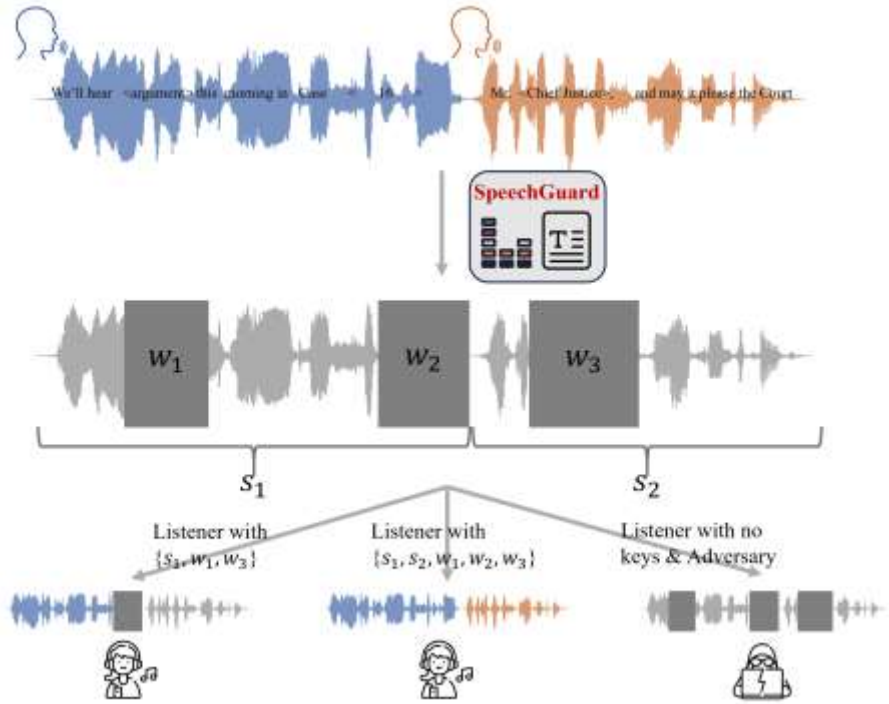
- Adversaries and Assumptions
  - Adversary capabilities:
    - Access to published protected audio.
    - Guess protection parameters and keys.
    - Use Automatic ASV to infer speaker identity.
    - Use Automatic ASR to infer protected content.
  - Both L_2 and L_3 may act as adversaries.

Threat Model

- Adversaries and Assumptions
  - Adversary capabilities:
    - Access to published protected audio.
    - Guess protection parameters and keys.
    - Use Automatic ASV to infer speaker identity.
    - Use Automatic ASR to infer protected content.
  - Both L_2 and L_3 may act as adversaries.
  - Cloud servers are honest-but-curious: no tampering, but may try to extract private info.

Solution Overview

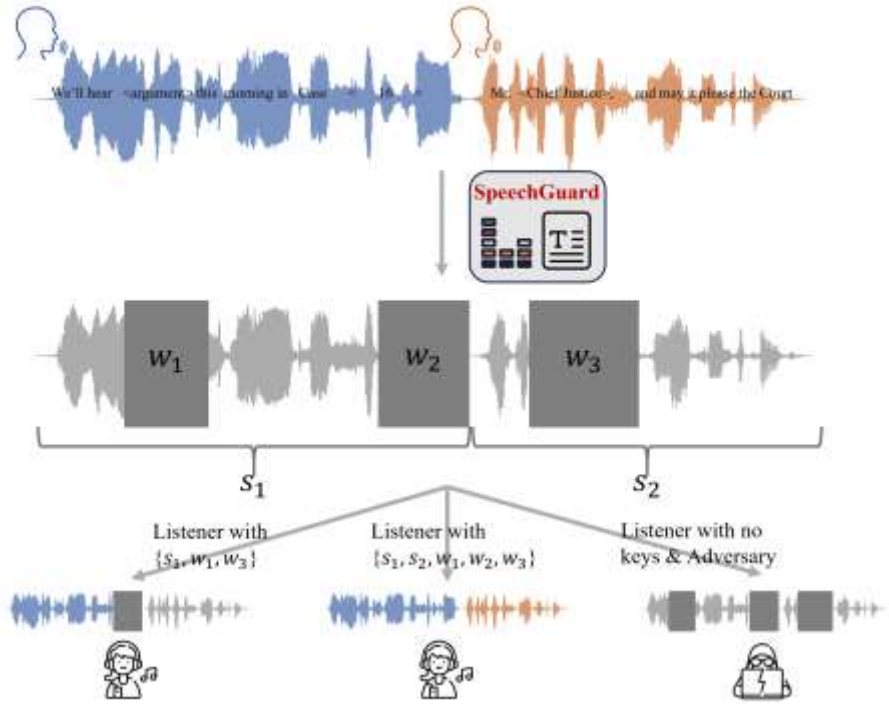
- Audio Owner Side
  - Reversible acoustic privacy protection: VTLN-based voice conversion → generates *warping parameters*.



Basic functions of SpeechGuard.

Solution Overview

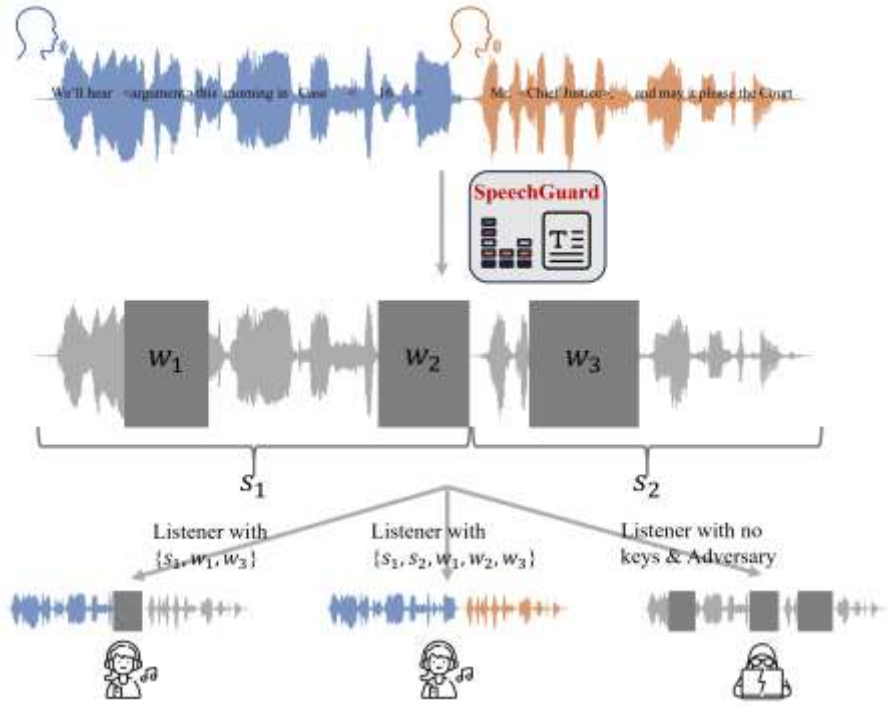
- Audio Owner Side
  - Reversible acoustic privacy protection: VTLN-based voice conversion → generates *warping parameters*.
  - Reversible content privacy protection: sensitive text detection & encryption → generates *encryption keys*.



Basic functions of SpeechGuard.

Solution Overview

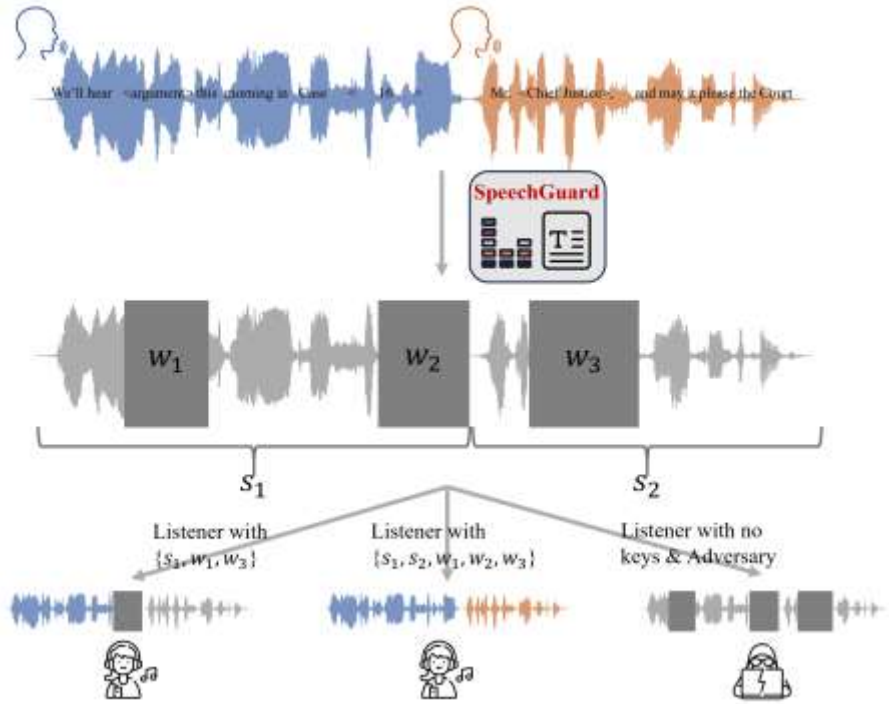
- Audio Owner Side
  - Reversible acoustic privacy protection: VTLN-based voice conversion → generates *warping parameters*.
  - Reversible content privacy protection: sensitive text detection & encryption → generates *encryption keys*.
  - Access control: distributes *warping parameters* and *encryption keys* according to permissions.



Basic functions of SpeechGuard.

Solution Overview

- Audio Owner Side
  - Reversible acoustic privacy protection: VTLN-based voice conversion → generates *warping parameters*.
  - Reversible content privacy protection: sensitive text detection & encryption → generates *encryption keys*.
  - Access control: distributes *warping parameters* and *encryption keys* according to permissions.
- Listener Side
  - Uses private keys to recover authorized acoustic and content information.



Basic functions of SpeechGuard.

Data Preprocessing

- Frame Splitting
  - Split the speech into fixed 36-ms frames.
  - Each frame is processed independently for fine-grained privacy control.

Data Preprocessing

- Frame Splitting
  - Split the speech into fixed 36-ms frames.
  - Each frame is processed independently for fine-grained privacy control.
- Segmentation
  - Group frames into voiced and unvoiced segments using a dynamic-threshold Voice Activity Detection (VAD).
  - Segments serve as the unit for subsequent voice conversion operations.

VTLN-Based Voice Conversion

- Basic Warping Function and Its Vulnerability
  - Piecewise Linear Function

$$p(\omega,\alpha)=\begin{cases}\alpha\omega, \text{if } \omega \le \omega_0 \\ \alpha\omega_0+\dfrac{\pi-\alpha\omega_0}{\pi-\omega_0}(\omega-\omega_0), \text{if } \omega > \omega_0\end{cases}$$

    Maps frequency ω using parameter α and turning point ω_0.
  - Invertibility
    - Knowing α allows recovery using inverse function $p^{-1}(\omega,\alpha)$.

D. Sundermann et al. "VTLN-based voice conversion." ISSPIT 2003.

VTLN-Based Voice Conversion

- Basic Warping Function and Its Vulnerability
  - Piecewise Linear Function

$$p(\omega,\alpha)=\begin{cases}\alpha\omega, \text{if } \omega \le \omega_0 \\ \alpha\omega_0+\dfrac{\pi-\alpha\omega_0}{\pi-\omega_0}(\omega-\omega_0), \text{if } \omega > \omega_0\end{cases}$$

    Maps frequency ω using parameter α and turning point ω_0.
  - Invertibility
    - Knowing α allows recovery using inverse function $p^{-1}(\omega,\alpha)$.
  - Vulnerability
    - Single parameter can be guessed, enabling reversal.
    - Reducing attack: approximate guesses still yield near-original voice.

D. Sundermann et al. "VTLN-based voice conversion." ISSPIT 2003.

VTLN-Based Voice Conversion

- Multi-Parameter Warping Function

$$m(\omega, \alpha, \beta) = \frac{\beta_i - \beta_{i-1}}{\alpha_i - \alpha_{i-1}}(\omega - \alpha_i) + \beta_i, \text{if } \alpha_{i-1} \leq \omega \leq \alpha_i$$

$$m^{-1}(\omega, \alpha, \beta) = m(\omega, \beta, \alpha)$$

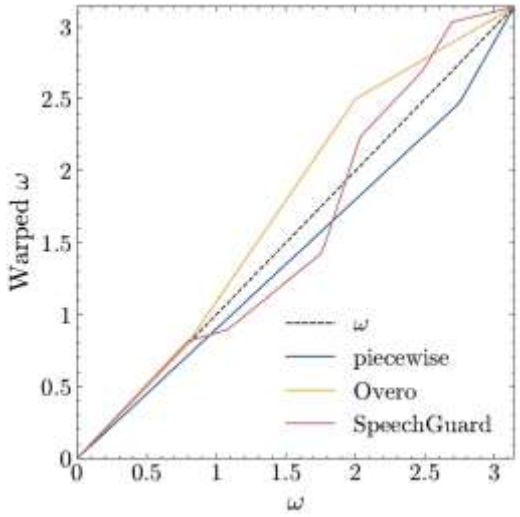
  - Randomly select the number of parameters n.
  - Generate parameter pairs $\alpha = \{\alpha_1, \dots, \alpha_n\}, \beta = \{\beta_1, \dots, \beta_n\}$.
  - Each segment uses a unique parameter set $s = \{\alpha, \beta\}$; parameters collected as $S = \{s_1, \dots, s_v\}$.

VTLN-Based Voice Conversion

- Multi-Parameter Warping Function

$$m(\omega, \alpha, \beta) = \frac{\beta_i - \beta_{i-1}}{\alpha_i - \alpha_{i-1}}(\omega - \alpha_i) + \beta_i, \text{if } \alpha_{i-1} \leq \omega \leq \alpha_i$$

$$m^{-1}(\omega, \alpha, \beta) = m(\omega, \beta, \alpha)$$

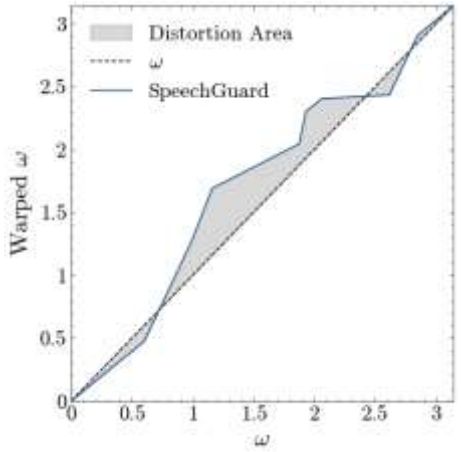


Multi-parameter warping provides stronger protection with more parameters, randomized set size, random parameter selection, and diverse distortion directions.

D. Sundermann et al. "VTLN-based voice conversion." ISSPIT 2003.
J. Lim et al. "Overo: Sharing private audio recordings." CCS 2022.

VTLN-Based Voice Conversion

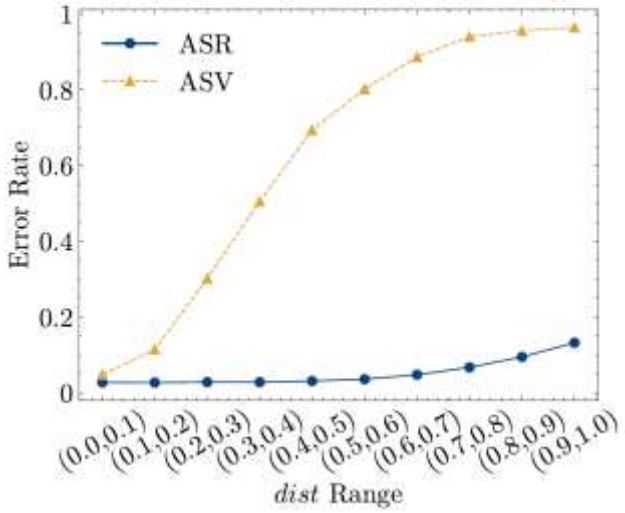
- Appropriate Range for Warping Parameters
  - Distortion strength $dist$
    - Defined as the area between the warping curve and the identity function.
    - Higher $dist$ → stronger distortion.



Distortion strength $dist$.

VTLN-Based Voice Conversion

- Appropriate Range for Warping Parameters
  - Goal: Balance privacy (ASV error ↑) and speech quality (ASR error ↓).
  - Appropriate range: $dist \in (0.5,0.6)$. Owners can adjust $dist$ depending on privacy needs.



ASR error rate rises slowly with distortion,
while ASV error rate spikes sharply.

Sensitive Text Detection & Encryption

- Goal: Protect content privacy by encrypting sensitive texts → encrypt all frames from start to end of the text.

Sensitive Text Detection & Encryption

- Goal: Protect content privacy by encrypting sensitive texts → encrypt all frames from start to end of the text.
- Detection
  - Automatic: use Speaker Diarization (SD) to obtain speaker and transcript segments, and Named Entity Recognition (NER) to detect sensitive patterns.
  - Manual: owners can adjust, add, or delete sensitive texts after recording to fix detection errors.

Sensitive Text Detection & Encryption

- Goal: Protect content privacy by encrypting sensitive texts → encrypt all frames from start to end of the text.
- Detection
  - Automatic: use Speaker Diarization (SD) to obtain speaker and transcript segments, and Named Entity Recognition (NER) to detect sensitive patterns.
  - Manual: owners can adjust, add, or delete sensitive texts after recording to fix detection errors.
- Encryption
  - Each sensitive text assigned a key; keys collected as $W = \{w_1, \dots, w_c\}$.

Sensitive Text Detection & Encryption

- Goal: Protect content privacy by encrypting sensitive texts → encrypt all frames from start to end of the text.
- Detection
  - Automatic: use Speaker Diarization (SD) to obtain speaker and transcript segments, and Named Entity Recognition (NER) to detect sensitive patterns.
  - Manual: owners can adjust, add, or delete sensitive texts after recording to fix detection errors.
- Encryption
  - Each sensitive text assigned a key; keys collected as $W = \{w_1, \dots, w_c\}$.
- Adaptation
  - Supports MP3 — only encrypts data frames, not headers.

Authorization & Privacy Recovery

- Frame-level reversible operations → different listeners see different privacy levels.

Authorization & Privacy Recovery

- Frame-level reversible operations → different listeners see different privacy levels.
- Access control
  - Acoustic privacy: warping parameters $S = \{s_1, \dots, s_v\}$.
  - Content privacy: encryption keys $W = \{w_1, \dots, w_c\}$.

Authorization & Privacy Recovery

- Frame-level reversible operations → different listeners see different privacy levels.
- Access control
  - Acoustic privacy: warping parameters $S = \{s_1, \dots, s_v\}$.
  - Content privacy: encryption keys $W = \{w_1, \dots, w_c\}$.
- Permission groups: Audio owner assigns subsets of S and W to listener groups for fine-grained control.

Authorization & Privacy Recovery

- Frame-level reversible operations → different listeners see different privacy levels.
- Access control
  - Acoustic privacy: warping parameters $S = \{s_1, \dots, s_v\}$.
  - Content privacy: encryption keys $W = \{w_1, \dots, w_c\}$.
- Permission groups: Audio owner assigns subsets of S and W to listener groups for fine-grained control.
- Secure distribution: Use Ciphertext-Policy Attribute-Based Encryption (CP-ABE) to deliver parameters/keys to authorized groups.

Evaluation

- Superior Acoustic Privacy Protection & Robustness
  - Against Reducing Attacks: SpeechGuard achieves the lowest attack success rate, 6.80% lower than the next best method, Overo. After an attack, the ASV Equal Error Rate (EER) increases by an average of 2.40%, further enhancing anonymity.
  - High Anonymity, Maintained Speech Quality: Protected speech demonstrates an average ASV EER of 33.48% (comparable to VoiceMask, the best baseline), while maintaining a low ASR Word Error Rate (WER) of 10.19%, ensuring good speech quality.
  - High-Fidelity Recovery: With authorization, speech achieves high-fidelity recovery, with ASV EER reduced to 7.07% and ASR WER to 9.90%, closely approaching the original speech.

Evaluation

- Unique Support for Recoverable & Customizable Content Privacy
  - Sensitive Content Confidentiality: The False Negative Rate (FNR) for encrypted sensitive words significantly increases to an average of 98.70%, ensuring sensitive information is unrecognizable.
  - Flexible MP3 Support: MP3 format shows slightly higher sensitive content protection performance (98.70% FNR) than that of WAV (97.38% FNR) and reduces file size to only 10% of WAV, optimizing storage and transmission.
  - Only Recoverable Solution: SpeechGuard is the only solution that allows encrypted sensitive text to be recovered to its original form, unlike other baseline methods (VoiceMask, Preech, Overo), which are irreversible.

Evaluation

- Practicality, Efficiency & User Satisfaction
  - Fine-Grained Access Control: Comprehensive evaluation validates the effectiveness of L_1 (full recovery), L_2 (partial recovery), and L_3 (no recovery) permission levels, enabling audio owners to finely customize access for different listeners.
  - Efficient Real-time Performance: The Real-time Coefficient (RTC) for the entire protection process is approximately 0.86, and for recovery, it's about 0.27, demonstrating efficient real-time processing capabilities.
  - High User Acceptance: User study results indicate that participants rated SpeechGuard 's manual annotation usability, protection effectiveness, recovery performance, and overall performance as "good" or "excellent."

Limitations and Future Work

- Voice conversion parameters validated only on three English datasets; to be expanded to other languages.
- Privacy and user experience depend on the accuracy of offline SD and NER models; further improvements needed.
- Sensitive text detection currently supports only specific words and patterns; will be enhanced for semantic-level detection using advanced NLP and contextual analysis.

Conclusion

- SpeechGuard is the first to unify privacy, recoverability, and customizability for speech.
- We design a multi-parameter, reversible warping function for voice conversion to achieve stronger anonymity and better resistance to reducing attacks.

Contact: jingmiao@mail.ustc.edu.cn