Supplimentary material

A   Visualization of Mask Generator

We have visualized some masks generated by the trained mask generator in Figure 11. It can be seen that the mask generator can dispatch suitable mask granularity to proper speech granularity to some extent. With more semantics utterance around, the mask becomes more meticulous, with the slices being distributed accordingly.
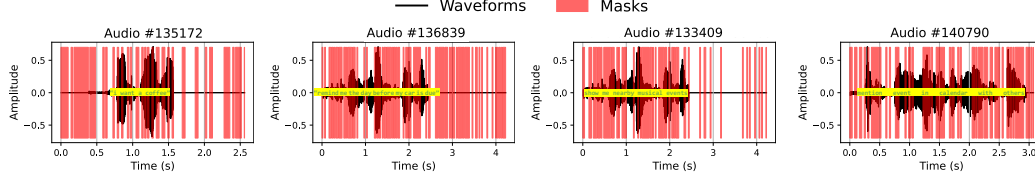
Figure 11: Illustration of the generated masks on audios selected randomly from SLURP. Local utterances are efficiently disrupted according to different transcripts patterns as highlighted within.

B   Additional Experiments

Active Inpainting Attacks.   We have implemented two active reconstruction adversaries and demonstrated our efficiency in defending against them. U-Net is a traditional inpainting model based on convolutional U-Net structure, commonly used in literature to reconstruct missing audio signals. We utilize the SLURP training set and their masked counterparts to train the inpainting model from scratch to reconstruct the missing audio. CQT-Diff is a neural diffusion model with an invertible Constant-Q Transform (CQT) to leverage pitch-equivariant symmetries, allowing it to effectively reconstruct audio without retraining. The reconstructed audio is sent to Whisper for automatic recognition. The visualizations of reconstructed waveforms are shown in Figure 12. The updated evaluation results under attacks are summarized in the table below.
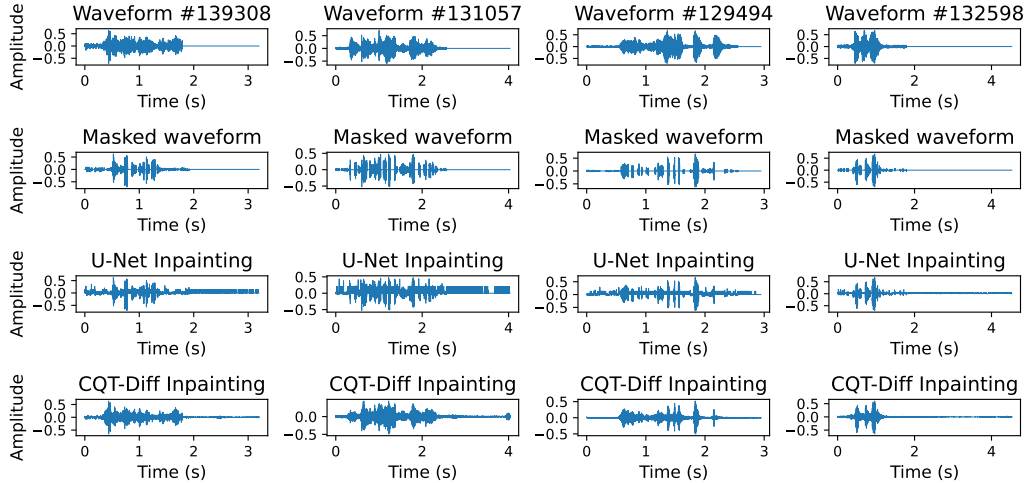
Figure 12: The reconstructed waveforms of different active inpainting attacks. Dataset: SLURP.

Detailed analysis of FSC dataset. We conducted further experiments on the Fluent Speech Commands (FSC) dataset, another widely used dataset for spoken language understanding research. The FSC dataset includes 97 speakers and 30,043 relevant utterances. We split the data, using 20% for testing and the remaining 80% for training. The results are shown in Table 2. The table shows that SILENCE achieves 99.1% SLU accuracy, with a 81.4% WER-ASR, outperforming all baselines. The results are consistent with the SLURP dataset, demonstrating the robustness of SILENCE across different datasets.

15

	AllOffloaded	VAE	PPSLU	Local	Random	SILENCE
ACC-SLU (%)	99.7	98.3	99.2	99.7	86.4	99.1
WER-ASR (%)	1.2	65.5	78.5	100	76.6	81.4

Table 2: Evaluation of privacy preservation and SLU performance on FSC dataset.

Integration with conventional SLU methods. We applied our algorithm to conventional modularized SLU models. The experimental results, shown in Table 3, demonstrate that when both the ASR and NLU modules are fine-tuned as required, the conventional modularized SLU model can recognize intent correctly when fed with masked audio. The detailed results are summarized in the table below:

	Plaintext	VAE	PPSLU	NLU only (Ours)	Decoupled SLU (Ours)	E2E SLU (Ours)
SLU-ACC (%)	87.2	72.5	74.5	12.6	89.1	81.1

Table 3: System performance on conventional modularized SLU.

Effect of mask granularity at various speech granularity. We included two more fine-grained speech understanding tasks: action and the combined intent (scenario_action) recognition. There are 18 different scenarios and 46 defined actions, resulting in 828 possible combinations for intend. As shown in Table 4, our method can recognize speech intent at different granularities. For example, we can correctly recognize 76.8% of the combined intent. In comparison, disentanglement-based methods need to re-entangle representations for different semantic granularities. Thus, the classifier used for scenario classification cannot be applied to other intents, and these methods are not designed to preserve the sensitive information within command audios. This emphasises a significant advantage of our approach, as it does not require retraining the model for different intent granularities.

	AllOffloaded	VAE	PPSLU	OnDevice	Ours
ACC-Scenario (%)	88.2	72.8	73.9	88.2	80.2
ACC-Action (%)	77.1	/	/	77.1	76.4
ACC-Intent (%)	83.3	/	/	83.3	76.8
WER-SLU (%)	14.7	/	/	100	68.6
WER-ASR (%)	12.3	69.3	75.3	100	68.1

Table 4: Comparison between Privacy-preservation and SLU performance at different speech granularities. '/' means not supported. Local leaks no words as nothing is uploaded.

16