Investigation of Voice Privacy Challenge Baseline B1 performance under different noise conditions

Saikeerthi Chirumamilla Hagadur[1], Ünal Ege Gaznepoglu[2], Nils Peters[2]

[1] *Friedrich-Alexander-Universität, 91058 Erlangen, Germany, Email: saikeerthi.chirumamilla@fau.de*

[2] *International Audio Laboratories Erlangen[*], 91058 Erlangen, Germany, Email: {ege.gaznepoglu, nils.peters}@audiolabs-erlangen.de*

Introduction

Previous years witnessed great progress in speech processing technologies. Some of these advancements, such as wide-scale adoption of voice assistants, and DeepFake generators capable of cloning speech with limited data, led to a need for privacy-preserving systems. Personal information such as demographics, or health conditions could be inferred from speech, even though the user's intention was not to share them in the first place.

While numerous approaches such as homomorphic encryption, deletion, and federated learning are imaginable, they are either not feasible, render the signal useless, or fail to protect the user's privacy [1]. For this reason, an approach called speaker anonymization has recently emerged. It aims to remove the speaker information from the speech signal, while preserving the necessary information, such as linguistic content or emotions.

Most publications in the field adopt the VoicePrivacy Challenge (VPC) framework, providing researchers with datasets, evaluation metrics, and baselines. VPC evaluation framework contains only clean speech recordings and does not reflect real-world use cases. While there has been investigations of the reproduction capabilities, e.g., as in [2], this was also limited to clean acoustic conditions. As a result, in this work, we investigate the robustness of the VPC Baseline B1 under different noise conditions. We focus on Gaussian noise and babble noise to contaminate the challenge datasets, perform anonymization, and finally compute and report the VPC evaluation metrics.

VoicePrivacy Challenge

VPC is the largest public event dedicated to speaker anonymization. The first edition was held in 2020, with the main aim of spearheading the development of privacy-preserving methods for speech. The VoicePrivacy Initiative has introduced various baselines, datasets, and evaluation metrics to create a level playing field.

Datasets

Table 1 provides a summary of the datasets utilized in this work. Detailed description of the corpora used is available in the evaluation plan [3].

Table 1: Speaker and utterance counts in the used datasets.

Purpose	Dataset	Female	Male	Utterances
Develop	LibriSpeech(Subset)	35	34	2321
Eval	LibriSpeech(Subset)	36	33	1934

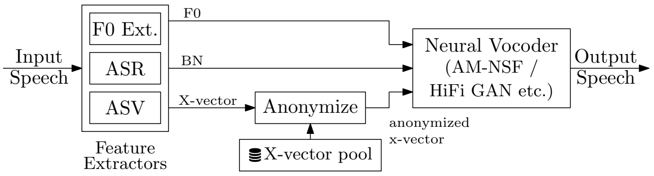


Figure 1: Speaker anonymization using X-vector and neural waveform models from [4]

Considered Anonymization System

We conduct our experiments on the system introduced in [4], for which an illustration is available in Figure 1. The system consists of semantically meaningful feature extractors, an anonymization block, and a speech synthesizer. Semantically meaningful features comprise of fundamental frequency (F0) representing the prosody of the speech, automated speech recognition (ASR) bottleneck features (BNs) for the linguistic content and X-vectors carrying the speaker identity information. The anonymization block generates a non-existing voice (called a pseudo-identity), reasonably different than the original speaker in a pseudorandom manner. The synthesis block combines the modified X-vector with the rest of the features and performs waveform synthesis.

Evaluation

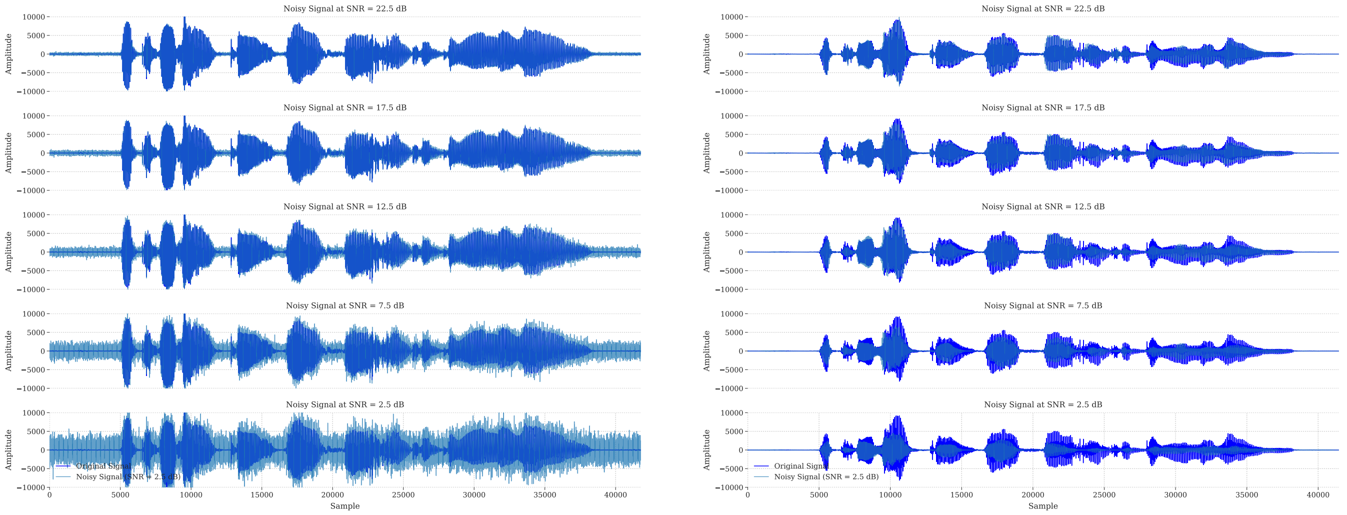
Attack models: Attack models outline the capabilities of the adversarial party that aims to create a privacy threat. In this paper, we consider three of the attack models adopted by the VPC framework, which are explained below. Figure 3 shows a summary of the considered attack models.

Unprotected (o-o): Serves as a benchmark, and measures how successful the automated speaker verification (ASV) system is at associating different unprocessed utterances from the same speaker.
Ignorant (o-a): Assumes the attacker has access to some unmodified speech signals and anonymized speech.
Lazy-Informed (a-a): In addition to the previous, assumes access to the anonymization system with different pseudo-random behavior (e.g., different random number generator seed).

Evaluation metrics: Two essential metrics to assess a speaker anonymization system's performance are the equal error rate (EER) and word error rate (WER).

Equal Error Rate: EER is the primary metric to mea-

[*] The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS.



(a) Original

(b) Anonymized

Figure 2: Subplots displaying comparison of original (a) and anonymized (b) signals with Gaussian noise at various SNR levels

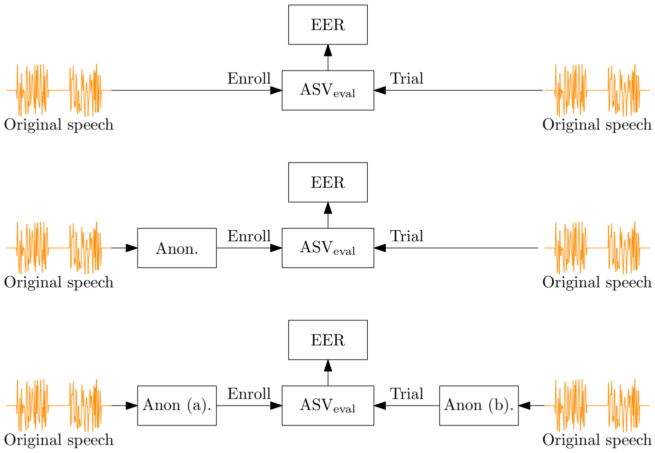


Figure 3: (Top) unprotected, (middle) ignorant, and (bottom) lazy-informed attack models

sure the privacy preservation success. An ASV system is utilized to see if the anonymized utterances could be associated with the original speaker. A higher EER value is desired, since we aim to reduce the linkability between the original speaker and the corresponding anonymized speech.

Word Error Rate: WER is the primary metric to measure the utility, by measuring the accuracy of an ASR system by comparing the transcriptions to the ground truth. Low WER indicates high accuracy in recognizing words, hence lower values imply a better performance.

Experiments

As outlined in the previous sections, the VPC framework performs evaluations in clean acoustic conditions. On the other hand, possible deployment scenarios, e.g., as a pre-processing step to the voice assistants, require robustness to different acoustically adverse conditions yet in the literature this has not been evaluated until now. To address this gap, we investigated the VPC B1's performance under two types of input speech degradations: White Gaussian Noise (WGN) and babble noise, because these acoustic conditions represent challenging scenarios commonly found in real-world environments. For both noise types, we will consider various noise strengths.

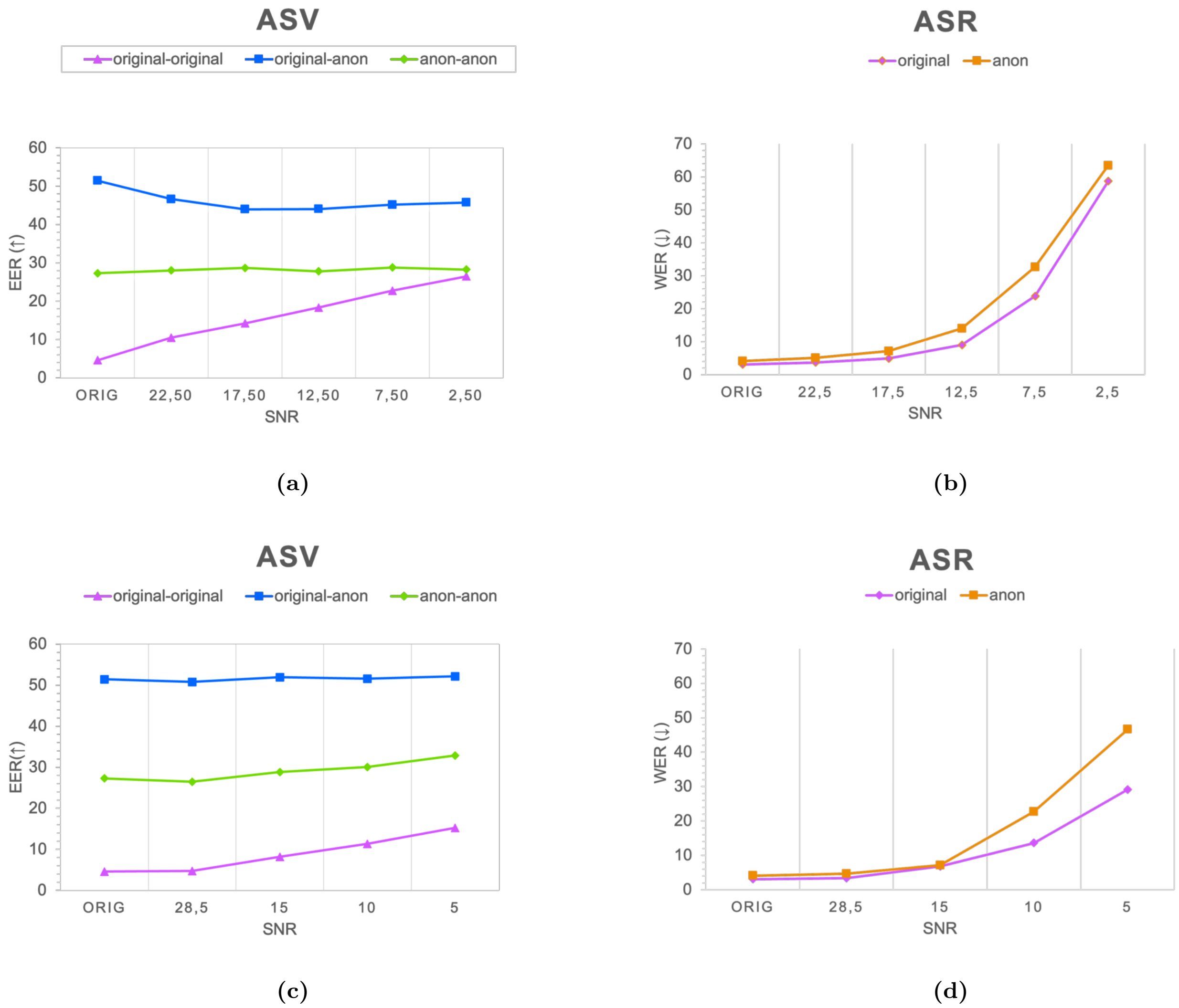
Gaussian Noise Gaussian noise is the signal noise that has a probability density function equal to the normal distribution and it is commonly used to simulate the sensor noise. We added Gaussian noise with varying SNRs to the speech data and ran the anonymization system on them. By looking at the visualizations of the waveforms produced by VPC B1 (see Fig. 2), the noise is not replicated at the anonymized output.

Babble Noise Babble noise is the presence of background speech of multiple people talking simultaneously, imitating possible acoustic conditions in closed living spaces. For example, the noise heard at restaurants, offices, or parties. To simulate such real-world scenarios, we have utilized Dinner Party Corpus (DiPCo) [5]. To create noisy utterances, we first made use of the available metadata to extract segments containing active speech. Afterwards, these segments are concatenated and for each utterance in our experiment set, we take a random slice, adjust the SNRs, and mix them additively, to create babble noise utterances.

Evaluation Results

Gaussian Noise Figure 4 depicts the results of ASV evaluation for (o-o), (o-a), and (a-a) attack models, along with ASR evaluation for original and anonymized utterances. The introduction of Gaussian noise showed limited impact on the ASV evaluation. EER values remained relatively stable, around 50% for (o-o) comparisons and approximately 30% for (a-a) comparisons. The mild decrease in (o-a) scores do not have a significant impact on the user privacy.

Regarding the ASR evaluation, it is observed for lower SNR values, WER values showed a significant rise.



(a)

(b)

(c)

(d)

Figure 4: ASV and ASR evaluation plots for Gaussian (a, b) and babble noise conditions (c, d)

Similarly, notable disparity is observed between the anonymized data and the input data, particularly evident at SNR levels of 12.5 dB and 7.5 dB. After listening to the voice samples, we concluded that the unvoiced phonemes are particularly susceptible to alterations under WGN.

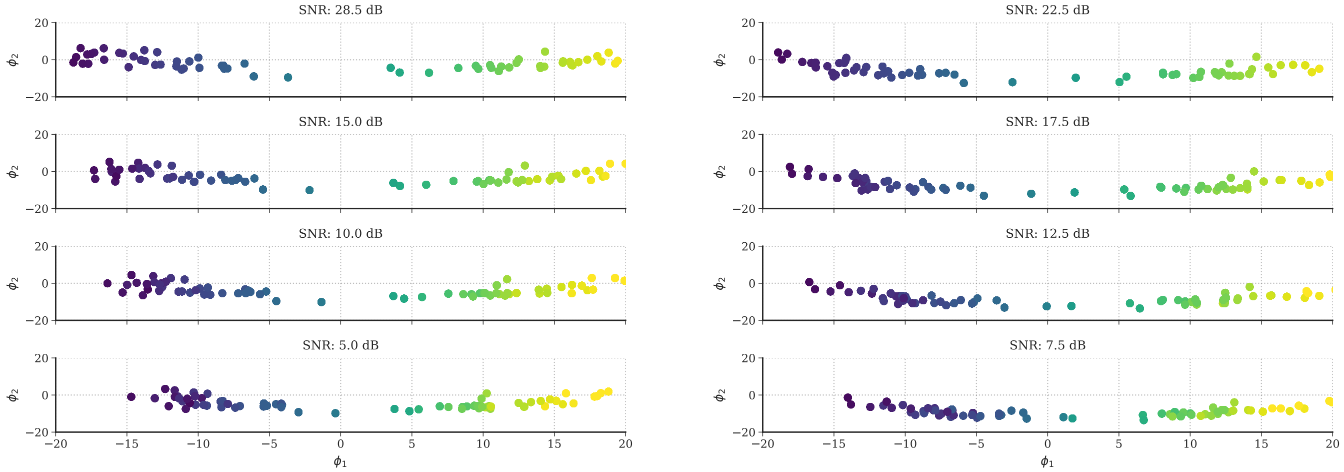
Babble Noise The evaluation of Automatic Speaker Verification ASV showed a mild impact on system performance, particularly observed in the (a-a) comparison. In this comparison, the EER experienced a moderate increase of 4% in absolute. The slope of (a-a) attack model is parallel to the (o-o) comparison, possibly indicating that the increase is caused by the presence of the second person's speech.

Discussion

Given input speech with considered noise types, the anonymization system under test successfully masked speaker identities, highlighting its effectiveness in privacy preservation. However, the linguistic content of the speech is affected by noise, particularly in WGN scenario. The pre-trained ASV evaluation system showed limited robustness to WGN noise, suggesting the necessity of including noisy data in the training set as well as some architectural modifications that improve the robustness of the evaluation system.

To gain further understanding of the ASV behavior, we trained a principal component analysis (PCA) model on the noise-free data, then visualized the first two principal components of the X-vectors for all noise levels in Figure . We observed that in noise-free setting, the first principal component is strongly correlated with the gender of the speaker, and we observed two clusters for female and male speakers. In the presence of WGN, the shape of the two clusters becomes morphed and their centers progress to the center. For babble noise, the shapes of the two clusters are largely undisturbed but the shift to the center is also visible. Based on these observations, we hypothesize that the x-vector extractor within the baseline would benefit from structural changes, e.g., as proposed by [6].

Future work Possible directions to explore in upcoming studies include retraining ASV evaluation systems with noisy data, exploring additional types of noise and degradation (e.g., reverberation), and investigating met-



(a) Gaussian noise

(b) Babble noise

Figure 5: (a) Projection of speaker-level X-vectors for different SNRs onto noiseless X-vector principal components space

rics for emotion preservation introduced in the VoicePrivacy Challenge 2024 [7] provide clear understandings of the observed behavior.

Conclusion

In conclusion, this study offers findings of our experiments on the VPC B1, to shed light on its behavior when input speech suffers from noise degradation. Anonymization performance, assessed by ASV-EER in the VPC context, indicates that the considered system would still preserve the privacy in the presence of noise. However, limited robustness of ASV evaluation system when presented with noisy speech signals is concerning, and hints the need for refinement and adaptation. By addressing these challenges and exploring potential enhancements, we can advance the development of privacy-preserving speech technologies that are better equipped to handle the complexities of real-world environments.

References

[1] N. Tomashenko *et al.*, “Introducing the VoicePrivacy Initiative,” in *Proc. Interspeech Conf.*, 2020.

[2] Ü. E. Gaznepoglu and N. Peters, “Evaluation of the Speech Resynthesis Capabilities of the VoicePrivacy Baseline B1,” in *Proc. 3rd Symp. on Security and Privacy in Speech Communication*, 2023.

[3] N. Tomashenko *et al.*, *2nd VoicePrivacy Challenge Evaluation Plan*, 2022.

[4] F. Fang *et al.*, “Speaker anonymization using x-vector and neural waveform models,” in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019.

[5] M. V. Segbroeck *et al.*, “DiPCo - Dinner Party Corpus,” in *Proc. Interspeech Conf.*, 2020.

[6] J. Thienpondt and K. Demuynck, *ECAPA2: A Hybrid Neural Network Architecture and Training Strategy for Robust Speaker Embeddings*, 2024.

[7] Pierre Champion *et al.*, *3rd VoicePrivacy Challenge Evaluation Plan*, 2024.