SRPOL voice hacking: methods of attacking speaker verification anonymization algorithms

*Equal contribution

Justyna Krzywdziak*
Samsung R&D Poland
Warsaw, Poland
j.krzywdziak@samsung.com

Ivan Ryzhankow*
Samsung R&D Poland
City, Country
i.ryzhankow@samsung.com

Szmajdziński Szymon*
Samsung R&D Poland
Warsaw, Poland
s.szmajdzins@samsung.com

Władysław Średniawa*
Samsung R&D Poland
Warsaw, Poland
w.sredniawa@samsung.com

Jakub Żak*
Samsung R&D Poland
Warsaw, Poland
jakub.zak@samsung.com

Adam Cieślak
Samsung R&D Poland
Warsaw, Poland
a.cieslak@samsung.com

Piotr Masztalski
Samsung R&D Poland
Warsaw, Poland
p.masztalski@samsung.com

I. Introduction

This document is a detailed description of the systems developed for the Attacker Challenge. We provide the background for each method used, confirming its selection, as well as the dataset and data pre-processing steps. As each anonymization system chosen for this challenge is unique and has its own strengths and weaknesses, we have developed separate attacker systems for each anonymization method. The II section consists of the method description used. Section III reveals used data set and preprocessing steps. In section IV you can find details about the methods used to develop systems for each anonymization technique, since for some of them we decided to use more than one method described in section II at the same time. In this section you can also find obtained results.

II. Methodologies

A. Knowledge Distillation

We decided to took advantage of the widely used knowledge distillation technique and used it to train the attacker system. The main purpose of using this method was to efficiently train the attacker model and let it keep the information about the original data, since it may contain some hidden important speaker-specific information. Our teacher and student model was ECAPA-TDNN [1] checkpoint from SpeechBrain [2] on Hugging Face. The teacher model was fed the original audio data while student model was fed the anonymized data. The loss was the sum of the Additive Angular Margin loss [3] calculated on anonymized data predictions from student model, and MSE loss between the student embedding of anonymized data and teacher embedding of original data multiplied by the distillation coefficient.

B. Audio Hashing

Perceptual hashing is a fingerprinting algorithm that produces a hash of various domains of data, including images. A perceptual hash is a locality-sensitive hash, which does not destroy the relation of features in the data. This is in contrast to cryptographic hashing, which is destructive to features. Perceptual hash functions provide the ability to find similar data by correlation of hashes [4]–[6]. In this approach, perceptual hashing was introduced to knowledge distillation. Instead of requiring from the student to replicate the teacher embeddings directly, the teacher embeddings were hashed, resulting in binary vectors of the same length. The student was then trained on those binary embeddings, with sigmoid activation at the head. The teacher model, student model and the loss calculation were the same as in the subsection II-A.

C. Data Mixing

Data mixing is an augmentation technique commonly used in image processing ML. The basic premise is to extract a fragment of one image (sample) and insert it in a random location of a second image (sample), replacing the pixel values completely [7]–[10]. For this task, an adjusted technique was used. The model (also ECAPA-TDNN) was first trained on personalized data. Then, each epoch, a percent of the anonimized sample was chosen and replaced the same location of personalized sample. The model was trained for another epoch on those mixed samples. With each epoch, the data received more anonimized fragments to replace the personalized ones. During the last epoch, the data was fully anonymized. The percent of data mixing is based on the number of epochs, i.e. the longer the training, the smaller the changes to data each epoch.

III. Datasets

The Librispeech dataset [11] is one of the most widely used and popular datasets in the field of speech recognition. It contains over 1000 hours of human speech recordings in English, derived from publicly available audiobook recordings. These recordings have been carefully processed and paired with transcriptions, making the dataset suitable for training various speech verification systems. To develop our models we used solely the Librispeech corpus. To use the original and anonymized data together we did the preprocessing stage that consisted of correct matching of chunked original-anonymized pairs, as the anonymization methods also affected the speed of speech. We examined the ratio of the length of the original recording to the anonymized recording and determined where the same fragment of speech would start and end for both recordings. These fragments were cut and the shorter of the fragments was padded with zeros at the beginning and end. This step ensured the unity of expression in both recordings so that we could exclude the possible focus of the model on differences that would come from the content of the recordings.

IV. Attacker systems for anonymization methods

A. B3

For this anonymization method we used data mixing method. In training process we set the amount of mel coefficients to 80, trained the model for 20 epochs with batch size 256. Additve Angular Margin loss was used with margin parameter set to 0.2 and scale set to 30. We used cyclical learning rate policy with base lr 0.000001, max lr 0.01 and 65000 step with Adam optimizer. Each epoch the amount of anonymized data increased by 5% till it reached 100% in the last epoch. We used input normalization with sentence norm_type. The audio was chunked and 3-seconds fragments were fed into the model.

TABLE I
Results for B3 method

	libri dev f	libri dev m	libri test m	libri test f
EER	32.238	29.222	32.118	27.618

B. B4

For this anonymization method we used data mixing, knowledge destillation and hashing methods. In training process we set the amount of mel coefficients to 80, trained the model for 50 epochs with batch size 256, although the most satisfying results were obtained in 12 epoch. Additive Angular Margin loss was used with margin parameter set to 0.2 and scale set to 30. We used cyclical learning rate policy with base lr 0.00000001, max lr 0.001 and 65000 step with Adam optimizer. Each epoch the amount of anonymized data increased by 2% till it reached 100% in the last epoch. We used input normalization with sentence norm_type. The audio was chunked and 3-seconds fragments were fed into the model.

TABLE II
Results for B4 method

	libri dev f	libri dev m	libri test m	libri test f
EER	34.658	30.153	27.026	29.397

C. B5

dest hashing For this anonymization method we used knowledge destillation and hashing methods. In training process we set the amount of mel coefficients to 80, trained the model for 50 epochs with batch size 256, although the most satisfying results were obtained in 12 epoch. Additve Angular Margin loss was used with margin parameter set to 0.2 and scale set to 30. We used cyclical learning rate policy with base lr 0.00000001, max lr 0.001 and 65000 step with Adam optimizer. We used input normalization with sentence norm_type. The audio was chunked and 3-seconds fragments were fed into the model.

TABLE III
Results for B5 method

	libri dev f	libri dev m	libri test m	libri test f
EER	32.804	30.312	30.136	30.954

D. T10-2

For this anonymization method we used data mixing, knowledge destillation and hashing methods. In training process we set the amount of mel coefficients to 80, trained the model for 50 epochs with batch size 256, although the most satisfying results were obtained in 22 epoch. Additive Angular Margin loss was used with margin parameter set to 0.2 and scale set to 30. We used cyclical learning rate policy with base lr 0.00000001, max lr 0.001 and 65000 step with Adam optimizer. Each epoch the amount of anonymized data increased by 2% till it reached 100% in the last epoch. We used input normalization with sentence norm_type. The audio was chunked and 3-seconds fragments were fed into the model.

TABLE IV
Results for T10-2 method

	libri dev f	libri dev m	libri test m	libri test f
EER	44.177	37.736	38.685	38.685

E. T12-5

For this anonymization method we used data mixing method. In training process we set the amount of mel coefficients to 80, trained the model for 20 epochs with batch size 256. Additive Angular Margin loss was used with margin parameter set to 0.2 and scale set to 30. We used cyclical learning rate policy with base lr 0.000001, max lr 0.01 and 65000 step with Adam optimizer. Each epoch the amount of anonymized data increased by 5% till it reached 100% in the last epoch. We used input normalization with sentence norm_type. The audio was chunked and 3-seconds fragments were fed into the model.

TABLE V
RESULTS FOR T12-5 METHOD

	libri dev f	libri dev m	libri test m	libri test f
EER	41.336	39.169	37.029	36.524

REFERENCES

[1] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.

[2] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[3] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[4] P. Samanta and S. Jain, "Analysis of perceptual hashing algorithms in image manipulation detection," *Procedia Computer Science*, vol. 185, pp. 203–212, 2021.

[5] S. McKeown and W. J. Buchanan, "Hamming distributions of popular perceptual hashing techniques," *Forensic Science International: Digital Investigation*, vol. 44, p. 301509, 2023.

[6] L. Du, A. T. Ho, and R. Cong, "Perceptual hashing for image authentication: A survey," *Signal Processing: Image Communication*, vol. 81, p. 115713, 2020.

[7] H. Naveed, S. Anwar, M. Hayat, K. Javed, and A. Mian, "Survey: Image mixing and deleting for data augmentation," *Engineering Applications of Artificial Intelligence*, vol. 131, p. 107791, 2024.

[8] D. Lewy and J. Mańdziuk, "An overview of mixing augmentation methods and augmentation strategies," *Artificial Intelligence Review*, vol. 56, no. 3, pp. 2111–2169, 2023.

[9] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "A comprehensive survey of image augmentation techniques for deep learning," *Pattern Recognition*, vol. 137, p. 109347, 2023.

[10] D. Sun, F. Dornaika, and J. Charafeddine, "Lcamix: Local-and-contour aware grid mixing based data augmentation for medical image segmentation," *Information Fusion*, vol. 110, p. 102484, 2024.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.