# Speaker De-identification System using Autoencoders and Adversarial Training

*Fernando M. Espinoza-Cuadros*[1,2], *Juan M. Perero-Codosero*[1,2], *Javier Antón-Martín*[1,2], *Luis A. Hernández-Gómez*[2]

[1]Sigma Technologies S.L.U., Madrid, Spain
[2]GAPS Signal Processing Applications Group, Universidad Politécnica de Madrid, Madrid, Spain
{fmespinoza,jmperero,janton}@sigma-ai.com, luisalfonso.hernandez@upm.es

## Abstract

The fast increase of web services and mobile apps, which collect personal data from users, increases the risk that their privacy may be severely compromised. In particular, the increasing variety of spoken language interfaces and voice assistants empowered by the vertiginous breakthroughs in Deep Learning are prompting important concerns in the European Union to preserve speech data privacy. For instance, an attacker can record speech from users and impersonate them to get access to systems requiring voice identification. Hacking speaker profiles from users is also possible by means of existing technology to extract speaker, linguistic (e.g., dialect) and paralinguistic features (e.g., age) from the speech signal. In order to mitigate these weaknesses, in this paper, we propose a speaker de-identification system based on adversarial training and autoencoders in order to suppress speaker, gender, and accent information from speech. Experimental results show that combining adversarial learning and autoencoders increase the equal error rate of a speaker verification system while preserving the intelligibility of the anonymized spoken content.

**Index Terms**: Speaker de-identification, Adversarial Training, Autoencoders, Adversarial Neural Networks

## 1. Introduction

Recent European privacy legislation, i.e., General Data Protection Regulation (GDPR), has limited some uses of the data in order to protect the personal information. According to the recent regulations, stored biometric data need to be unlinkable, irreversible, and renewable [1]. This is the case of speech, which is considered personal information by itself. The main reason is that speech contains extra information apart from spoken contents. Furthermore, the emergent use of voice assistants has made that spoken commands are used by applications to carry out different actions. Sometimes it is necessary to collect some speech to improve and adapt the assistant's models to the user's speech. In this case, an attacker could have access to sensitive user's data (e.g., not only several utterances, but also some speaker profiles, such as age, gender. that can be easily obtained from these utterances). Thus, the objective is privacy preservation, suppressing critical speaker information from speech.

In order to preserve speaker privacy, some solutions have been proposed. Cryptography-based solutions involve a large complexity and computational overhead. Instead, anonymization is more flexible allowing also the removal of personally identifiable information within a speech signal. Since there is not a formal definition of anonymization (de-identification), VoicePrivacy initiative is defining metrics, protocols, and a benchmark on common datasets. Thus, privacy preservation solutions will be developed to anonymize the speech, but maintaining intelligibility and naturalness [2].

The speaker de-identification task aims to suppress the speaker identity, which might be represented in the linguistic content of the speaker's speech [3, 4] and spectral and excitation features of the speech signal [5, 6]. Previous studies in speaker de-identification area are very limited. Most of them are based on voice transformation (VT) systems [6, 7], and phoneme recognition followed by speech synthesis from the phoneme sequence [8]. In [7], authors proposed an improved VT-based approach to enable the speaker to be de-identified by voice transformation from a pool of pre-trained VT models. In [8], authors proposed the de-identification of the real speaker by averaging a set of x-vectors from a pool of pre-trained x-vectors and select the most dissimilarity x-vector. This final x-vector along with the sequence of diphones and fundamental frequency (F0) synthesize the anonymized speech.

Domain-Adversarial training (DAT) [9] has been applied to improve Automatic Speech Recognition (ASR) performance by learning features invariant to various conditions, such as acoustic variabilities [10, 11], accented speech [12], and inter-speaker feature variability [13, 14, 15]. Similarly, these techniques have been applied for speaker privacy protection. Speaker privacy protection in [16] uses adversarial training to generate representations that perform well in ASR while hiding speaker identity. Along the same lines, in [17] speaker-invariant training is carried out via reconstruction network in addition to the DNN acoustic model and trained jointly via adversarial training. Following the same approach, in this paper, we propose a speaker de-identification method based on the combination of adversarial training and autoencoders in order to generate speaker-invariant features as well as to other speaker characteristics (i.e., gender and accent). The rest of the paper is organized as follows. In Section 2, we describe the proposed speaker de-identification system based on x-vector anonymization. Section 3 explains the experimental setup under the VoicePrivacy 2020 Challenge [2]. Results are presented and discussed in Section 4. Finally, conclusions and future work are given in Section 5.

## 2. Speaker de-identification system

The speech signal contains different sources of variability. The speaker-dependent variability has been used to develop speaker characterization systems (e.g., age [18], gender [19], pathologies [20, 21], among others). In other cases, as in Automatic Speech Recognition (ASR), the speaker variability together with the acoustic environment variabilities (e.g., noise, channels, etc.) are considered undesired sources of variability. This has led to the proposal of different techniques to remove the ef-

fect of speaker [12, 13, 14, 15] and the noise conditions in [10], [11] to improve the ASR accuracy. Similarly, in [17], the use of a reconstruction network an a DNN acoustic model is jointly optimized through adversarial multi-task learning to generate speaker-invariant features.

Following a similar approach, we propose a speaker de-identification method using DAT [9] and Autoencoders. Our method does not start from scratch, but it is based on the Baseline-1 anonymization system proposed in VoicePrivacy 2020 Challenge [2], from now on referred as the baseline. It consists of three main parts: 1) feature extraction, 2) x-vector anonymization, and 3) speech synthesis. In this work we address part 2) of the baseline x-vector anonymization. In our approach, Adversarial Training and Autoencoders are proposed to remove information related to the speaker's characteristics in the anonymized x-vector while preserving an acceptable ASR performance.

### 2.1. Autoencoder-Adversarial Network

Based on speaker de-identification methods using adversarial networks presented in [14, 16, 17], we propose a speaker-characteristics-invariant approach based on an Autoencoder-Adversarial Network (AAN). In the proposed ANN architecture (see Fig. 1), an encoder-decoder autoencoder branch tries to reconstruct the input x-vector while in adversarial branches we try to mitigate speaker characteristics, such as gender, accent and speaker identity. From this approach, we aim to hide the speaker identity when reconstructing the x-vector by means of the autoencoder but making the latent or encoded representation invariant to the domain of speaker characteristics by using a Domain Adversarial Neural Network (DANN) [9]. For this adversarial architecture, we are now given a training dataset denoted as $\{x_i, y_i, z_{gi}, z_{ai}, z_{si}\}_N^{i=1}$, where $x_i$ and $y_i$ are the original and reconstructed x-vector, respectively, and $z_{gi}$, $z_{ai}$, $z_{si}$ are the different domain classes as gender, accent, and speaker identity respectively of the i-th data point. We denote the $\theta_e$ and $\theta_d$ the parameters of the latent representation and decoder of the autoencoder respectively, and by $\theta_g$, $\theta_a$ and $\theta_s$ the parameters of the gender, accent and speaker identity of the adversarial branches respectively. The objective function for the autoencoder $L_{au}$ and adversarial branches $L_z$ are defined as

$$L_{au}(\theta_e, \theta_d) = -\sum_{i=i}^{N} \log P(y_i|x_i; \theta_e, \theta_d) \quad (1)$$

$$L_z(\theta_e, \theta_g, \theta_a, \theta_s) = -\sum_{k\in\{g,a,s\}} \sum_i \log P(z_{si}|x_i; \theta_e, \theta_k) \quad (2)$$

Thus, our model is trained by optimizing the following min-max objective:

$$\min_{\theta_e,\theta_d} \max_{\theta_e,\theta_g,\theta_a,\theta_s} L_{au}(\theta_e, \theta_d) - \lambda L_z(\theta_e, \theta_g, \theta_a, \theta_s), \quad (3)$$

where $\lambda$ is a trade-off parameter between the autoencoder objective and the adversarial objectives, which goal is to remove the speaker characteristics via backpropagation by means of the Gradient Reversal Layer (GRL) [9] in each adversarial branch.

### 2.2. X-vector anonymization approaches

Based on the proposed framework, we evaluate two different approaches for x-vector anonymization. Both approaches use the x-vector extractor from the baseline system. In the first approach, the AAN described before, is used as x-vector anonymizer. That is, the x-vectors extracted from the baseline

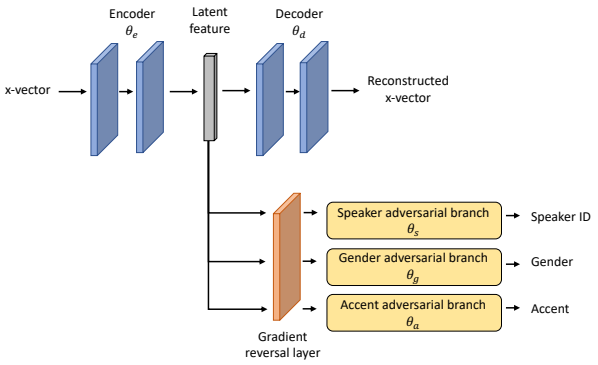


Figure 1: *Autoencoder-Adversarial Network (AAN) architecture.*

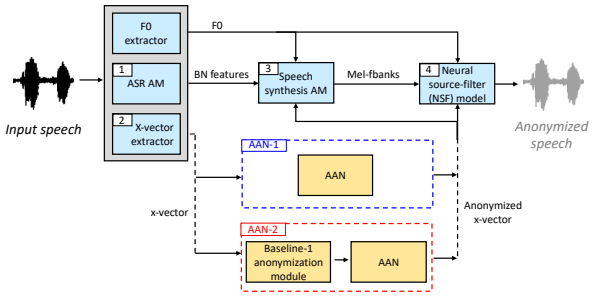


Figure 2: *X-vector anonymization approaches (i.e. AAN-1 and AAN-2) based on Autoencoder-Adversarial Networks. Adapted from [2].*

are used as input to the autoencoder that generates as output the pseudo-speaker x-vector, as shown in Fig. 2 (AAN-1 label). In the second approach, we transform the pseudo-speaker x-vector generated by the baseline. As it can be seen in Fig. 2 (AAN-2 label), the anonymized x-vector is used as input to the autoencoder that generates as output a new anonymized x-vector.

### 2.3. ANN Network architecture and training

The enconder-decoder autoencoder model consists of 4 dense layers of size 512 with *tanh* activation functions, trained using Mean Square Error (MSE) loss function. Each adversarial branch consists of a dense layer of size 128 and ReLU activation followed by a *softmax* output layer, which output dimension corresponds to the number of classes on each adversarial feature. For the feature *accent* the output dimension is 30, i.e., the number of accents in training dataset, for *speaker* is 1251, i.e., speakers and for *gender* is 2, i.e., female and male. Cross-entropy loss function was used on the adversarial training.

For AAN training and testing we used VoxCeleb-1 [22], which contains approximately 330 hours of recordings from 1251 speakers. It also contains gender and accent information for each speaker. For AAN training, a closed-set speaker identification task was performed for the speaker adversarial branch. We assign 10 utterances per speaker to validation and test. The remaining utterances were used for training. For the rest of the adversarial branches, gender and accent labels were used. To select the optimal trade-off-parameter $\lambda$, several values were tested running the anonymization task of the VoicePrivacy 2020 Challenge. The best results were achieved for $\lambda = 8$.

# 3. Experimental setup

## 3.1. Dataset

The proposed anonymization system, for both ANN-1 and ANN-2 approaches, was evaluated accordingly to VoicePrivacy 2020 Challenge using LibriSpeech [23] and VCTK [24] datasets for both ASR (intelligibility) and Automatic Speaker Verification (ASV) (anonymization) evaluation tasks. A detailed description of both datasets used in the challenge can be found in [2].

## 3.2. Evaluation system and metrics

The anonymization system performance was evaluated by means of the assessment of the speaker verifiability and the ability of the anonymization system to preserve the intelligibility of the anonymized spoken content, which is carried out by pretrained $\text{ASV}_{eval}$ and $\text{ASR}_{eval}$ models provided by the VoicePrivacy Challenge. The metrics for both ASV and ASR tasks evaluation are Equal Error Rate (EER) and Log-likelihood-ratio cost function (Cllr), and Word error rate (WER), respectively. A detailed description of these metrics can be found in [2].

# 4. Results and discussion

In this section, we present the results for ASV (Tables 5 and 6) and ASR (Tables 2 and 3) tasks of the Challenge for both ANN-1 and ANN-2 anonymization approaches. We also compare our results to those from the baseline (Tables 1 and 4).

Overall, when compared to the baseline system, our results show that both ANN proposals increase speaker de-identification while providing similar intelligibility of the anonymized spoken content. However, the increase in speaker de-identification is only observed for both-sides anonymization condition (a-enroll, a-trial). That is when comparing to the baseline (Table 4), ANN results for the original-enroll and anonymized-trial condition (a-enroll, o-trial) in Tables 5 and 6 show, in the worst case, a decrease in speaker de-identification performance, in terms of EER, of approximately 9% and 8% for AAN-1 and AAN-2 respectively. In contrast, in the both-sides anonymization condition (a-enroll, a-trial), both approaches achieve better speaker de-identification results. For the best-case, we can observe an increase in performance over the baseline, in terms of EER, of approximately 9% and 10% for AAN-1 and AAN-2 respectively. Results in Table 5 and 6, also show that for both best and worst- case scenarios, similar results are obtained in both ANN approaches.

Differences in performance between the ANNs approaches and the baseline for different anonymization conditions can be related to the performance of the proposed x-vector anonymization methods. For the o-enroll, a-trial condition, in the x-vector anonymization baseline, there is a high chance that the anonymized x-vector is very different from the original one as it corresponds to an average of farthest x-vectors from the original. Whereas in the AAN-1 approach, as the autoencoder aims to reconstruct the original x-vector while suppressing the speakers' characteristics via adversarial training, there is a less probability that the reconstructed x-vector in the anonymized trials will be very far from the original. In the ANN-2 approach, we could expect that the anonymization performance should overcome the baseline due to the addition of variability to the anonymized x-vector. Nevertheless, that is not the case since the ANN training it is not optimized to reconstruct the anonymized x-vector. In contrast, in the both-sides anonymization condition (a-enroll, a-trial), we believe that the baseline system may have a high chance that the anonymized x-vectors in the enrollment and the trial fall in the same region since the anonymized x-vectors are selected from the farthest x-vectors from the original utterances on both sides, which belong to the same speaker. Thus, there is a chance that the anonymized x-vectors on both sides might be closed to each other. Differently from that, in our approach, the use of adversarial training for AAN introduces and additional variability to the reconstructed x-vector that can lead to the observed increase in speaker de-identification.

Table 1: *ASR results for **Baseline** for development and test data (o-original, a-anonymized speech).*

| # | Dev. set | WER, % | | Data | Test set | WER, % | |
|---|---|---|---|---|---|---|---|
| | | $\text{LM}_s$ | $\text{LM}_l$ | | | $\text{LM}_s$ | $\text{LM}_l$ |
| 1 | libri_dev | 5.25 | 3.83 | o | libri_test | 5.55 | 4.15 |
| 2 | libri_dev | 8.76 | 6.39 | a | libri_test | 9.15 | 6.73 |
| 3 | vctk_dev | 14.00 | 10.79 | o | vctk_test | 16.39 | 12.82 |
| 4 | vctk_dev | 18.92 | 15.38 | a | vctk_test | 18.88 | 15.23 |

Table 2: *ASR results for **AAN-1** for development and test data (a-anonymized speech).*

| # | Dev. set | WER, % | | Data | Test set | WER, % | |
|---|---|---|---|---|---|---|---|
| | | $\text{LM}_s$ | $\text{LM}_l$ | | | $\text{LM}_s$ | $\text{LM}_l$ |
| 1 | libri_dev | 9.22 | 6.75 | a | libri_test | 9.24 | 6.74 |
| 2 | vctk_dev | 18.67 | 15.20 | a | vctk_test | 19.09 | 15.16 |

Table 3: *ASR results for **AAN-2** for development and test data (a-anonymized speech).*

| # | Dev. set | WER, % | | Data | Test set | WER, % | |
|---|---|---|---|---|---|---|---|
| | | $\text{LM}_s$ | $\text{LM}_l$ | | | $\text{LM}_s$ | $\text{LM}_l$ |
| 1 | libri_dev | 9.28 | 6.76 | a | libri_test | 9.37 | 6.85 |
| 2 | vctk_dev | 18.69 | 15.25 | a | vctk_test | 19.04 | 15.21 |

Finally, as stated before, results for ANN-1 and ANN-2 in the ASR task (Tables 2 and 3) show a performance in intelligibility of the anonymized spoken content similar to that of the baseline system (Table 1).

# 5. Conclusions and future work

In this work, we present two methods for x-vector anonymization. These methods are integrated and evaluated under the Baseline-1 anonymization system proposed in VoicePrivacy 2020 Challenge. Both methods rely on an Autoencoder-Adversarial Network that tries to reconstruct x-vectors but intending to alleviate, through adversarial branches, the information of speaker characteristics in order to hide the speaker identity to a greater extent. Our experimental results show that though similar results to the baseline were achieved, when testing on both sides anonymization condition (i.e., training and testing with anonymized speech) our system outperforms the baseline. Those results foster to keep researching in adversarial training techniques, as well as the use of generative models to generate speaker-invariant features. Our future research will address the application of this framework to both the encoded speech content and the prosodic features looking for a better anonymization of the speech waveform suppressing the speaker information but preserving the spoken content.

Table 4: *ASV results for **Baseline** for development and test data (o-original, a-anonymized speech; **Gen** denotes speaker gender: **f**-female, **m**-male).*

| # | Dev. set | EER, % | $C_{llr}^{min}$ | $C_{llr}$ | Enroll | Trial | Gen | Test set | EER, % | $C_{llr}^{min}$ | $C_{llr}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | libri_dev | 8.665 | 0.304 | 42.857 | o | o | f | libri_test | 7.664 | 0.183 | 26.793 |
| 2 | libri_dev | 50.140 | 0.996 | 144.112 | o | a | f | libri_test | 47.260 | 0.995 | 151.822 |
| 3 | libri_dev | 36.790 | 0.894 | 16.345 | a | a | f | libri_test | 32.120 | 0.839 | 16.270 |
| 4 | libri_dev | 1.242 | 0.034 | 14.250 | o | o | m | libri_test | 1.114 | 0.041 | 15.303 |
| 5 | libri_dev | 57.760 | 0.999 | 168.988 | o | a | m | libri_test | 52.120 | 0.999 | 166.658 |
| 6 | libri_dev | 34.160 | 0.867 | 24.715 | a | a | m | libri_test | 36.750 | 0.903 | 33.928 |
| 7 | vctk_dev_com | 2.616 | 0.088 | 0.868 | o | o | f | vctk_test_com | 2.890 | 0.091 | 0.866 |
| 8 | vctk_dev_com | 49.710 | 0.995 | 172.049 | o | a | f | vctk_test_com | 48.270 | 0.994 | 162.531 |
| 9 | vctk_dev_com | 27.910 | 0.741 | 7.205 | a | a | f | vctk_test_com | 31.210 | 0.830 | 9.015 |
| 10 | vctk_dev_com | 1.425 | 0.050 | 1.559 | o | o | m | vctk_test_com | 1.130 | 0.036 | 1.041 |
| 11 | vctk_dev_com | 54.990 | 0.999 | 192.924 | o | a | m | vctk_test_com | 53.390 | 1.000 | 190.136 |
| 12 | vctk_dev_com | 33.330 | 0.840 | 23.891 | a | a | m | vctk_test_com | 31.070 | 0.835 | 21.680 |
| 13 | vctk_dev_dif | 2.864 | 0.100 | 1.134 | o | o | f | vctk_test_dif | 4.887 | 0.169 | 1.495 |
| 14 | vctk_dev_dif | 49.970 | 0.989 | 166.027 | o | a | f | vctk_test_dif | 48.050 | 0.998 | 146.929 |
| 15 | vctk_dev_dif | 26.110 | 0.760 | 8.414 | a | a | f | vctk_test_dif | 31.740 | 0.847 | 11.527 |
| 16 | vctk_dev_dif | 1.439 | 0.052 | 1.158 | o | o | m | vctk_test_dif | 2.067 | 0.072 | 1.817 |
| 17 | vctk_dev_dif | 53.950 | 1.000 | 167.511 | o | a | m | vctk_test_dif | 53.850 | 1.000 | 167.824 |
| 18 | vctk_dev_dif | 30.920 | 0.839 | 23.797 | a | a | m | vctk_test_dif | 30.940 | 0.834 | 23.842 |

Table 5: *ASV results for **AAN-1** for development and test data (o-original, a-anonymized speech; **Gen** denotes speaker gender: **f**-female, **m**-male).*

| # | Dev. set | EER, % | $C_{llr}^{min}$ | $C_{llr}$ | Enroll | Trial | Gen | Test set | EER, % | $C_{llr}^{min}$ | $C_{llr}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | libri_dev | 44.320 | 0.974 | 171.463 | o | a | f | libri_test | 43.980 | 0.972 | 168.557 |
| 2 | libri_dev | 39.630 | 0.921 | 22.336 | a | a | f | libri_test | 34.850 | 0.886 | 27.144 |
| 3 | libri_dev | 49.840 | 0.989 | 153.223 | o | a | m | libri_test | 45.430 | 0.980 | 155.451 |
| 4 | libri_dev | 43.480 | 0.964 | 36.897 | a | a | m | libri_test | 46.100 | 0.979 | 47.663 |
| 5 | vctk_dev_com | 50.580 | 0.976 | 183.167 | o | a | f | vctk_test_com | 47.110 | 0.982 | 171.678 |
| 6 | vctk_dev_com | 29.650 | 0.802 | 14.289 | a | a | f | vctk_test_com | 37.570 | 0.913 | 17.304 |
| 7 | vctk_dev_com | 47.860 | 0.985 | 171.920 | o | a | m | vctk_test_com | 44.920 | 0.984 | 172.326 |
| 8 | vctk_dev_com | 38.180 | 0.921 | 30.378 | a | a | m | vctk_test_com | 37.290 | 0.927 | 30.642 |
| 9 | vctk_dev_dif | 49.860 | 0.957 | 177.802 | o | a | f | vctk_test_dif | 48.350 | 0.997 | 155.964 |
| 10 | vctk_dev_dif | 30.430 | 0.828 | 14.852 | a | a | f | vctk_test_dif | 34.000 | 0.881 | 21.306 |
| 11 | vctk_dev_dif | 44.760 | 0.988 | 151.207 | o | a | m | vctk_test_dif | 48.160 | 0.996 | 157.427 |
| 12 | vctk_dev_dif | 33.800 | 0.882 | 30.406 | a | a | m | vctk_test_dif | 39.040 | 0.947 | 33.550 |

Table 6: *ASV results for **AAN-2** for development and test data (o-original, a-anonymized speech; **Gen** denotes speaker gender: **f**-female, **m**-male).*

| # | Dev. set | EER, % | $C_{llr}^{min}$ | $C_{llr}$ | Enroll | Trial | Gen | Test set | EER, % | $C_{llr}^{min}$ | $C_{llr}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | libri_dev | 45.880 | 0.981 | 171.212 | o | a | f | libri_test | 44.890 | 0.980 | 166.823 |
| 2 | libri_dev | 39.630 | 0.924 | 23.006 | a | a | f | libri_test | 35.770 | 0.898 | 27.617 |
| 3 | libri_dev | 50.000 | 0.992 | 153.313 | o | a | m | libri_test | 45.880 | 0.985 | 155.770 |
| 4 | libri_dev | 43.940 | 0.968 | 38.383 | a | a | m | libri_test | 46.770 | 0.979 | 48.320 |
| 5 | vctk_dev_com | 50.870 | 0.981 | 183.799 | o | a | f | vctk_test_com | 46.820 | 0.983 | 172.018 |
| 6 | vctk_dev_com | 29.070 | 0.815 | 15.106 | a | a | f | vctk_test_com | 39.020 | 0.917 | 17.719 |
| 7 | vctk_dev_com | 47.580 | 0.986 | 172.169 | o | a | m | vctk_test_com | 45.480 | 0.987 | 172.479 |
| 8 | vctk_dev_com | 39.320 | 0.936 | 31.763 | a | a | m | vctk_test_com | 38.700 | 0.943 | 32.110 |
| 9 | vctk_dev_dif | 50.480 | 0.963 | 179.226 | o | a | f | vctk_test_dif | 49.280 | 0.996 | 156.662 |
| 10 | vctk_dev_dif | 31.050 | 0.833 | 15.638 | a | a | f | vctk_test_dif | 34.830 | 0.895 | 21.842 |
| 11 | vctk_dev_dif | 45.310 | 0.991 | 151.928 | o | a | m | vctk_test_dif | 48.390 | 0.996 | 157.517 |
| 12 | vctk_dev_dif | 34.690 | 0.897 | 31.013 | a | a | m | vctk_test_dif | 39.610 | 0.955 | 33.957 |

# 6. References


[1] ISO/IEC JTC1 SC27 Security Techniques, "ISO/IEC 24745:2011. Information Technology-Security Techniques-Biometric Information Protection," International Organization for Standardization, Tech. Rep., 2011. [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso-iec:24745:ed-1:v1:en

[2] N. Tomashenko, B. Mohan, L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "The VoicePrivacy 2020 Challenge Evaluation Plan," pp. 1–17, 2020. [Online]. Available: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf

[3] S. H. K. Parthasarathi, M. Magimai-Doss, H. Bourlard, and D. Gatica-Perez, "Evaluating the robustness of privacy-sensitive audio features for speech detection in personal audio log scenarios," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2010, pp. 4474–4477.

[4] S. H. K. Parthasarathi, H. Bourlard, and D. Gatica-Perez, "Wordless sounds: Robust speaker diarization using privacy-preserving audio representations," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 1, pp. 83–96, 2013.

[5] K. Hashimoto, J. Yamagishi, and I. Echizen, "Privacy-preserving sound to degrade automatic speaker verification performance," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2016-May. Institute of Electrical and Electronics Engineers Inc., may 2016, pp. 5500–5504.

[6] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Speaker de-identification via voice transformation," in *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009*, 2009, pp. 529–533.

[7] M. Pobar and I. Ipšić, "Online speaker de-identification using voice transformation," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014 - Proceedings*. IEEE Computer Society, 2014, pp. 1264–1267.

[8] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker Anonymization Using X-vector and Neural Waveform Models," in *10th ISCA Speech Synthesis Workshop*. ISCA: ISCA, sep 2019, pp. 155–160. [Online]. Available: http://www.isca-speech.org/archive/SSW_2019/abstracts/SSW10_P_2-4.html

[9] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," *Advances in Computer Vision and Pattern Recognition*, no. 9783319583464, pp. 189–209, may 2015. [Online]. Available: http://arxiv.org/abs/1505.07818

[10] D. Serdyuk, K. Audhkhasi, P. Brakel, B. Ramabhadran, S. Thomas, and Y. Bengio, "Invariant Representations for Noisy Speech Recognition," *arXiv preprint arXiv:1612.01928*, nov 2016. [Online]. Available: http://arxiv.org/abs/1612.01928

[11] Y. Shinohara, "Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-Sept. International Speech and Communication Association, sep 2016, pp. 2369–2372. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2016/abstracts/0879.html

[12] S. Sun, C.-F. Yeh, M.-Y. Hwang, M. Ostendorf, and L. Xie, "Domain Adversarial Training for Accented Speech Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2018, pp. 4854–4858. [Online]. Available: https://ieeexplore.ieee.org/document/8462663/

[13] Y. Adi, N. Zeghidour, R. Collobert, N. Usunier, V. Liptchinsky, and G. Synnaeve, "To Reverse the Gradient or Not: An Empirical Comparison of Adversarial and Multi-task Learning in Speech Recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2019-May. Institute of Electrical and Electronics Engineers Inc., may 2019, pp. 3742–3746.

[14] Z. Meng, J. Li, Z. Chen, Y. Zhao, V. Mazalov, Y. Gang, and B. H. Juang, "Speaker-Invariant Training Via Adversarial Learning," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April. Institute of Electrical and Electronics Engineers Inc., sep 2018, pp. 5969–5973.

[15] T. Tsuchiya, N. Tawara, T. Ogawa, and T. Kobayashi, "Speaker Invariant Feature Extraction for Zero-Resource Languages with Adversarial Learning," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April. Institute of Electrical and Electronics Engineers Inc., sep 2018, pp. 2381–2385.

[16] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-Preserving Adversarial Representation Learning in ASR: Reality or Illusion?" *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, pp. 3700–3704, nov 2019. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2415

[17] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English Conversational Telephone Speech Recognition by Humans and Machines," in *Interspeech 2017*, vol. 2017-August. ISCA: ISCA, aug 2017, pp. 132–136. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0405.html

[18] P. Ghahremani, P. S. Nidadavolu, N. Chen, J. Villalba, D. Povey, S. Khudanpur, and N. Dehak, "End-to-end Deep Neural Network Age Estimation," in *Interspeech 2018*. ISCA: ISCA, sep 2018, pp. 277–281. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2018/abstracts/2015.html

[19] M. Kotti and C. Kotropoulos, "Gender classification in two Emotional Speech databases," in *Proceedings - International Conference on Pattern Recognition*, 2008.

[20] C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Pathological speech detection using x-vector embeddings," *arXiv preprint arXiv:2003.00864*, 2020.

[21] J. M. Perero-Codosero, F. Espinoza-Cuadros, J. Anton-Martin, M. A. Barbero-Alvarez, and L. A. Hernandez, "Modeling Obstructive Sleep Apnea voices using Deep Neural Network Embeddings and Domain-Adversarial Training," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8926341

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Interspeech 2017*. ISCA: ISCA, aug 2017, pp. 2616–2620. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0950.html

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2015-August. Institute of Electrical and Electronics Engineers Inc., aug 2015, pp. 5206–5210.

[24] V. Christophe, Y. Junichi, and M. Kirsten, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *The Centre for Speech Technology Research (CSTR)*, 2016.