PRIVACY-AWARE FEATURE EXTRACTION FOR GENDER DISCRIMINATION VERSUS SPEAKER IDENTIFICATION

Alexandru Nelus and Rainer Martin

Ruhr-Universität Bochum, Institute of Communication Acoustics, Bochum, Germany
Email: {alexandru.nelus,rainer.martin}@rub.de

ABSTRACT

This paper introduces a deep neural network based feature extraction scheme that aims to improve the trade-off between utility and privacy in speaker classification tasks. In the proposed scenario we develop a feature representation that helps to maximize the performance of a gender classifier while minimizing additional speaker identity information. Our approach is to use variational information feature extraction that allows for gender discrimination (utility) but minimizes the information level of the features, thus discouraging speaker identification adversarial attacks (privacy). We analyze the model's loss function and the budget scaling factor used to control the balance of utility vs. privacy. It is experimentally shown that the proposed method reduces privacy risks without significantly deprecating utility and that it also generalizes well to new speaker contexts.

***Index Terms*—** Privacy, utility, adversarial attack, variational information, mutual information, speaker classification

1. INTRODUCTION

The ubiquitous use of portable smart devices has also exposed us to an abundance of sensors such as microphones and cameras. Besides the obvious benefits there are also high privacy risks involved, especially when these sensor are connected in a (ad-hoc) sensor network. Evidently, it would be optimal if the sensors were designed to handle the desired task with maximum performance (*utility*) and at the same time minimize task-extraneous information (*privacy*). Considering that utility and privacy are competing goals and that they can not be simultaneously maximized, we like to explore the trade-off relation between utility and privacy in a speaker classification task.

A good illustrator of this concept is a scenario where a small office environment is host to a wireless acoustic sensor network (WASN) with distributed microphones and a processing scheme [1], [2] that performs deep neural network (DNN) based feature extraction at node level. The extracted feature representation is sent across the network to a DNN-based sink node which performs e.g. a speaker gender discrimination task. Without additional measures, the feature representation extracted for the sole purpose of gender discrimination also carries a significant amount of speaker-dependent data. Therefore, we propose to employ privacy-aware variational information feature extraction which produces a lossy, mutual information (MI) based information minimization and analyze its effects on the balance between the accuracy of gender discrimination (utility) and speaker identification (privacy). Due to the intrinsic relation between both classification tasks, this scenario is challenging and will not have a trivial solution.

The remainder of this paper is organized as follows: We first discuss the relation to prior work, we then describe the privacy-aware feature extraction model, followed by a description of the neural network architecture used, after which we detail the experimental layout and the results, finalizing with conclusions and ideas for future work.

2. RELATION TO PRIOR WORK

The topic of privacy-aware feature extraction was previously investigated by the authors in [1], where generative adversarial feature extraction was used to control the trade-off between gender discrimination and speaker identification. Although efficient, this method does not lead to a generalized information minimization technique due to its dependency on a specific attacker configuration. Therefore, the purpose of this paper is to introduce a more general approach.

The proposed solution is inspired by variational information autoencoders [3] where the encoding variable is a compact data representation and where a re-parametrization trick [4] is used to allow stochastic sampling during backpropagation. Our choice of MI as a regularization criterion is supported by works like [5] and [6], where it is successfully used to increase network performance and robustness against adversarial attacks in the testing domain. As far as the authors are aware, at the time of writing this paper, there is no previous investigation on using variational information networks against adversarial feature-interception attacks in WASNs.

3. DEFENDER VERSUS ATTACKER

3.1. Model description

We start with the model proposed in [1] where the concepts of *defender* and *attacker* are introduced and adapt this to the proposed scenario as shown in Fig. 1.

The defender consists of a feature extraction block f which transforms the low level feature set X into the high level feature set Z. The latter is then passed to the multilayer perceptron (MLP) based gender discriminator g with the weights and biases parameters Φ_g which in turn estimates the gender class labels' probabilities $P(\Gamma)$.

The feature extractor block f is composed of a convolutional neural network (CNN) based structure c with weights and biases parameters Φ_c which uses as input the low level feature set X. The CNN's output is concomitantly passed to the dense layers μ and σ which have the respective weights and biases parameters Φ_μ and Φ_σ. The output of the σ layer is multiplied with samples from a ζ-dimensional standard normal distribution $\mathcal{N}(0, I)$ and the result is then added to the output of the μ layer, creating thus the high level

This work has been supported by DFG under contract no. Ma 1769/6-1 within the Research Unit FOR 2457 (Acoustic Sensor Networks).

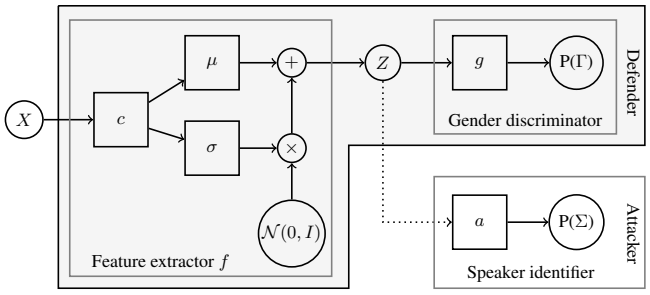

Fig. 1. Flow chart of privacy-aware feature extraction for gender discrimination vs. speaker identification.

feature set Z. The motivation behind this stochastic encoding is explained in the following section.

The attacker, which consists of an MLP-based classifier a with weights and biases parameters Φ_a, intercepts the high level feature set Z with which it estimate the speaker labels' probabilities $P(\Sigma)$.

3.2. Training the defender

In our example, the objective of the defender is to develop a feature extraction process that leads to good gender discrimination accuracy but which will result in bad classification accuracy if intercepted by an attacker for a more privacy invasive task such as speaker identification.

The first part of the objective can be formulated as minimizing the cross-entropy between the gender labels' true $P(\Gamma^t)$ and estimated $P(\Gamma)$ probability distributions as shown:

$$\min_{\Phi_c,\Phi_\mu,\Phi_\sigma} \mathbb{E}_{\Gamma^t\sim p(\Gamma^t)}[-\log p(\Gamma)]. \tag{1}$$

The second part of the objective can be addressed by minimizing the information in the high level feature set Z, thus rendering it as useless as possible to classification tasks. A good information regularization criterion is the mutual information $I(X;Z)$ between the input and output feature sets[5], [6]. Estimating this quantity is computationally challenging. A more practical solution is to find an MI upper bound $I_{max}(X;Z) \geq I(X;Z)$ and use this bound in the optimization process.

For this, we introduce the entropy-based MI formulation:

$$\begin{aligned} I(X;Z) &= H(Z) - H(Z|X) \\ &= -\int p(z)\log p(z)dz + \int p(x,z)\log p(z|x)dxdz. \end{aligned} \tag{2}$$

Recently, a stochastic encoding mechanism was introduced [4] in order to obtain an analytical expression of $H(Z|X)$. This is the reason why we construct a normal-distributed encoding variable $z = \mu(c(x)) \cdot \epsilon + \sigma(c(x))$, where $\epsilon \sim \mathcal{N}(0, I)$. In this way we force the conditional distribution of z given the input variable x to follow a Gaussian distribution:

$$p(z|x) = \mathcal{N}(\mu(c(x)), \sigma(c(x))). \tag{3}$$

Moreover, the stochastic sampling from $p(z|x)$ during backpropagation can be efficiently performed by updating the Φ_μ and Φ_σ parameters of layers μ and σ. This is referred to as the re-parametrization trick [4].

We are now left with finding an analytical expression for $H(Z)$. In this regard we introduce a variational distribution $q(z)$, which for simplicity we assume to be Gaussian $\mathcal{N}(0, I)$, as also suggested by [5]. Using the Kullback-Leibler divergence's property of always being positive [7] we obtain an upper bound for $H(Z)$:

$$\begin{aligned} KL(p(z)||q(z)) \geq 0 \Rightarrow \int p(z)\log\frac{p(z)}{q(z)}dz \geq 0 \\ \Rightarrow -\int p(z)\log p(z)dz \leq -\int p(z)\log q(z)dz. \end{aligned} \tag{4}$$

Using 2 and 4 we upperbound $I(X;Z)$ as:

$$\begin{aligned} I(X;Z) &\leq -\int p(z)\log q(z)dz + \int p(x,z)\log p(z|x)dxdz \\ &= -\int p(x,z)\log q(z)dxdz + \int p(x,z)\log p(z|x)dxdz \\ &= \int p(x,z)\log\frac{p(z|x)}{q(z)}dxdz = KL(p(z|x)||q(z)), \end{aligned} \tag{5}$$

where KL is the Kullback-Leibler distance between the conditional $p(z|x)$ and variational $q(z)$ distributions. We can now define the MI upper bound $I_{max}(X;Z) \geq I(X;Z)$ as:

$$I_{max}(X;Z) = KL(p(z|x)||q(z)). \tag{6}$$

In this way we can reduce the impact that the low level feature representation X has on the high level feature representation Z.

According to [8] and given 3 and that $q(z) = \mathcal{N}(0, I)$ we get:

$$I_{max}(X;Z) = \frac{1}{2}\Big(\mathrm{tr}(\Sigma_z) + \mu_z^\top\mu_z - \log\det(\Sigma_z) - \zeta\Big), \tag{7}$$

where $\Sigma_z = \mathrm{diag}(\sigma(c(x))^2)$ and $\mu_z = \mu(c(x))$.

Considering the earlier mentioned defender's competing goals of offering good gender discrimination accuracy, expressed by 1 while at the same time reducing task-extraneous information by minimizing $I_{max}(X;Z)$, the defender's loss function to be minimized can be formulated as:

$$\min_{\Phi_c,\Phi_\mu,\Phi_\sigma} \mathbb{E}_{\Gamma^t\sim p(\Gamma^t)}[-\log p(\Gamma)] + \beta I_{max}(X;Z). \tag{8}$$

Similarly to [1], [9] and [6] a *budget scaling* factor β is used to control how much gender discrimination accuracy we wish to renounce in favor of a more compressed high level feature representation.

3.3. Training the attacker

We use the feature extractor f to extract the high level feature representation Z which is then used by the attacker. The goal of the attacker is to perform speaker identification as best as possible using the intercepted feature set Z. This is done by minimizing the cross-entropy between the speaker labels' true $P(\Sigma^t)$ and estimated $P(\Sigma)$ probability distributions:

$$\min_{\Phi_a} \mathbb{E}_{\Sigma^t\sim p(\Sigma^t)}[-\log p(\Sigma)]. \tag{9}$$

4. NETWORK CONFIGURATION

4.1. Low level features extractor

We use the log mel-band energy (LMBE) representation of the signal $x_s(t)$ as the low level feature input for the neural network based feature extractor f. After applying a short-time Fourier transform

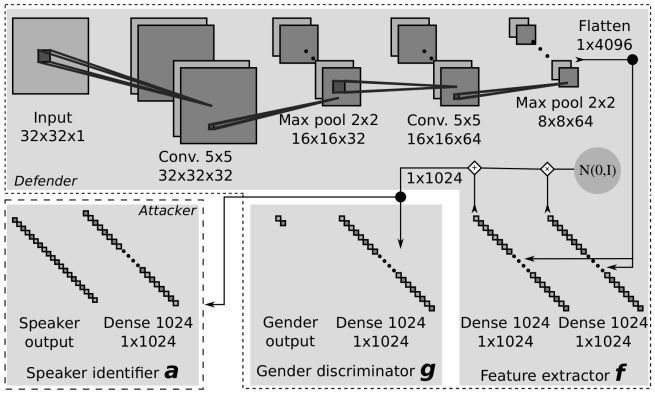

Fig. 2. Network architecture for privacy-aware variational information feature extraction.

(STFT) $X_{\text{stft}}(\kappa, b)$ with window length L_1 and step R_1 to $x_s(t)$, where κ and b denote the frequency bin and time frame index, respectively, we map the squared-magnitude spectrum onto the Mel scale [10], resulting in the Mel-spectrum $X_{\text{mel}}(k', b)$, where $k' = 0, 1, \ldots, K' - 1$ is the index of the Mel scale frequency bin. The LMBE features are then obtained by taking the logarithm of the absolute Mel-spectrum and keeping the first K'' coefficients:

$$X_{\text{lmbe}}(k', b) = \log |X_{\text{mel}}(k', b)|. \tag{10}$$

4.2. High level features extractor

The architecture of the feature extractor f is shown in Fig. 2. The CNN-based c block consists of two convolutional layers of sizes $32 \times K''$ and 16×16, each containing 32 and respectively 64 kernels of size 5×5 and rectified linear unit (ReLu) activation functions. Each layer is followed by a max-pooling layer of stride $S_1 = S_2 = 2$ and filter size 2×2. The extractor takes in a high resolution sample stream X_{lmbe} of the form $\left[\frac{T}{R_1}\right] \times K''$, where T is the signal's time length. The extractor's output is stacked in the form of $\left[\frac{T}{R_1 \times 32}\right] \times 4096$ and passed to the dense layers μ and σ, each containing ζ neurons. The output of layer σ is multiplied with samples from a ζ-dimensional standard normal distribution $\mathcal{N}(0, I)$ and added to the output of layer μ. The resulting high level feature representation has the form $\left[\frac{T}{R_1 \times 32}\right] \times \zeta$, where each feature vector is responsible for a receptive field of length $32R_1$ s.

4.3. Gender discriminator and speaker identifier

We perform gender discrimination and speaker identification for each resulting high level feature vector Z by using the MLP architectures g and respectively a presented in Fig. 2. Both MLP architectures consist of 1024 fully connected nodes that use ReLu activation functions and a final layer of two respectively S_t output nodes, on which we apply a softmax function. For training we employ the Adam optimizer [11] with a learning rate of 0.0001 and we also use a dropout rate of 0.4 [12].

<table>
<tr><td colspan="2">20 speakers WSJ</td><td colspan="2">80 speakers WSJ
240 speakers TIMIT</td></tr>
<tr><td>80 %
data/
speaker</td><td>20 %
data/
speaker</td><td colspan="2">Gender testing set WSJ
Gender testing set TIMIT</td></tr>
<tr><td rowspan="4">Training
set WSJ</td><td rowspan="4">Evaluation
set WSJ</td><td colspan="2">Speaker testing set WSJ:
4 batches × 20 speakers
Speaker testing set TIMIT:
20 batches × 20 speakers</td></tr>
<tr><td>80 %
data/speaker</td><td>20 %
data/speaker</td></tr>
<tr><td>Training
subset</td><td>Evaluation
subset</td></tr>
</table>

Table 1. Division of audio data into training, evaluation and testing sets, along with corresponding subsets.

5. EXPERIMENTS

5.1. Database and settings

The database contains 100 speakers from the WSJ corpus [13], of which 50 are male and 50 are female, with an average of 872 seconds (142 utterances) of audio per speaker. We select $S_t/2$ male and respectively female speakers, and randomly split every individual's audio data into training (80%) and evaluation (20%) sets (WSJ). The data from the remaining $100 - S_t$ speakers is used for the *gender testing* set (WSJ). The $100 - S_t$ speakers are also divided into subgroups of S_t, and every speaker's audio data is randomly split into *speaker training* (80%) and *speaker evaluation* (20%) subsets, together forming the *speaker testing* set (WSJ).

The database is supplemented with 420 speakers from the TIMIT corpus [14], of which 290 are male and 130 are female, with an average of 31 seconds (10 utterances) of audio per speaker, thus forming the gender testing set TIMIT. Additionally the 420 speakers are also divided into subgroups of S_t, and every speaker's audio data is randomly split into speaker training (80%) and speaker evaluation (20%) subsets, together forming the speaker testing set TIMIT. This structure is also described in Table 1, for $S_t = 20$ speakers.

We propose to use utterance level *accuracy* as a performance measure for both gender and speaker classification:

$$accuracy = \frac{\textit{no. of correctly classified utterances}}{\textit{total no. of utterances}}, \tag{11}$$

where the utterance's class label is assigned using a majority decision on the class labels of the utterance's feature frames.

The values of the system's parameters are: $S_t = 20$, $L_1 = 0.026$ s, $R_1 = 0.013$ s, $K'' = 32$, $\zeta = 1024$.

5.2. Feature extraction and classification

We first train the defender part of our model for an empirically selected number of 5000 iterations and a mini-batch size of 300 samples. Training is done on the training set WSJ and evaluation on the evaluation set WSJ. The accuracy of the latter is depicted in Fig. 3 under the label "Gender eval. WSJ". We then test the gender discrimination accuracy of the already trained model on the gender testing set WSJ and on the gender testing set TIMIT. The results are depicted in Fig. 3 under the labels "Gender test WSJ" and "Gender test TIMIT" respectively. This procedure is applied for the systematically varied values of the budget scaling factor β.

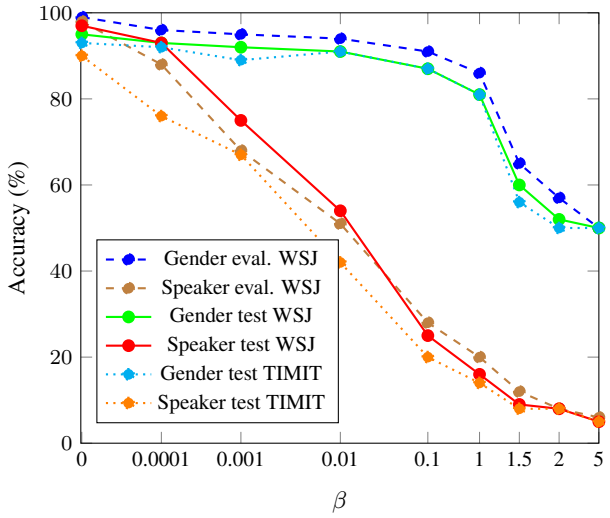

Fig. 3. The influence of the budget scaling factor β on gender discrimination and speaker identification accuracy using the WSJ and TIMIT data sets.

We next train the attacker part of our model by concatenating the previously trained feature extractor f with the attacker architecture detailed in Fig. 2. Training is performed on the training set WSJ for an empirically selected number of 5000 iterations and a mini-batch size of 300 samples. Evaluation is performed on the evaluation set WSJ. The accuracy of the latter is depicted in Fig. 3 under the label "Speaker eval. WSJ". We apply the same method to the speaker testing sets WSJ and TIMIT, where training is performed on the training subsets and evaluation on the evaluation subsets. The speaker identification accuracy results are depicted in Fig. 3 under the labels "Speaker test WSJ" and "Speaker test TIMIT" respectively. This procedure is also applied for the systematically varied values of the budget scaling factor β.

5.3. Discussion

Our first observation is that for $\beta = 0$, meaning that no mutual information regularization is used, the high level feature set Z which is extracted for the sole purpose of performing gender discrimination also carries a significant amount of speaker-dependent data, resulting in high speaker identification accuracy even for the smaller TIMIT data set.

As soon as β is increased, thus increasing the emphasis on the variational information minimization during the feature extraction process, a steep deprecation of speaker identification can be observed. For $\beta \leq 0.1$ the gender discrimination is not significantly affected, only dropping by a mean accuracy of 6%, while speaker identification drops by 70%. For larger values of β the deprecation of speaker identification continues but at a higher gender discrimination expense until for $\beta = 5$, both tasks report minimum performance, similar to random guessing. Moreover, deeper MLP speaker identifier architectures have also been used in trying to simulate a more powerful attacker but no significant deviation was observed.

6. CONCLUSIONS AND FUTURE WORK

We have empirically demonstrated that variational information feature extraction can be successfully employed to reduce the amount of task-extraneous information that DNN-extracted features inadvertently carry and thus strengthen their robustness against adversarial feature-interception attacks. For this, a privacy-aware network configuration along with a general loss function were proposed and a budget scaling factor was introduced and analyzed.

The two competing tasks and the databases were chosen as to best depict the proposed concept. In future works we aim to broaden the range of utility-based acoustic classification tasks and the range of privacy-invasive attackers. This will also consider non-parametric mutual information estimation.

7. REFERENCES

[1] Alexandru Nelus and Rainer Martin, "Gender discrimination versus speaker identification through privacy-aware adversarial feature extraction," in *Speech Communication; 13. ITG Symposium; Proceedings of*. VDE, 2018.

[2] Janek Ebbers, Alexandru Nelus, Rainer Martin, and Reinhold Haeb-Umbach, "Evaluation of modulation-MFCC features and DNN classification for acoustic event detection," in *Deutsche Jahrestagung fur Akustik (DAGA)*, 2018.

[3] Naftali Tishby, Fernando C Pereira, and William Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[4] Diederik P Kingma and Max Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[5] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[6] Yan Zhang, Mete Ozay, Zhun Sun, and Takayuki Okatani, "Information potential auto-encoders," *arXiv preprint arXiv:1706.04635*, 2017.

[7] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*, Wiley-Interscience, New York, NY, USA, 2006.

[8] John Duchi, "Derivations for linear algebra and optimization," vol. 3, chapter 9. Berkeley, California, 2007.

[9] Yusuke Iwasawa, Kotaro Nakayama, Ikuko Eguchi Yairi, and Yutaka Matsuo, "Privacy issues regarding the application of DNNs to activity-recognition using wearables and its countermeasures by use of adversarial training," in *Proceedings of the IJCAI*, 2017, pp. 1930–1936.

[10] Sadaoki Furui, *Digital speech processing: synthesis, and recognition*, CRC Press, 2000.

[11] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[12] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[13] Douglas B Paul and Janet M Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the Workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[14] Victor Zue, Stephanie Seneff, and James Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.