RUHR-UNIVERSITÄT BOCHUM

Department of Electrical Engineering and Information Technology
Institute of Communication Acoustics

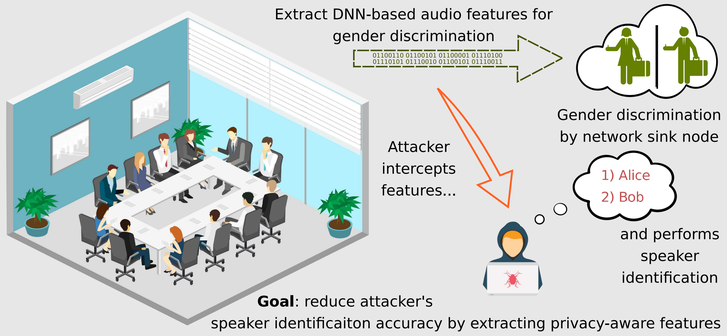
RUB

Privacy-Aware Feature Extraction For Gender Discrimination Versus Speaker Identification

Alexandru Nelus and Rainer Martin

Introduction

- Tackle **privacy** risks encountered in **Acoustic Sensor Network** applications
- Illustrate concept with a smart office and challenging competing goals scenario
- Balance competing goals: **utility** (gender discrimination) & **privacy** (speaker identification)



Goal: reduce attacker's speaker identificaiton accuracy by extracting privacy-aware features

Defender vs. Attacker

- Previous work [1]:
  - traditional feature representation carries significant speaker-dependent data
  - adversarial feature extraction successfully used but depends on attacker configuration
- More general approach: **privacy-aware variational information** feature extraction:
  - inspired by variational information autoencoders [2] which use information minimization
  - the encoding variable is a compact stochastic feature representation
  - the proposed system is described in Fig. 1

Train defender

- Z should lead to good gender discrimination accuracy while reducing task-extraneous data:

$$\min_{\Phi_c,\Phi_\mu,\Phi_\sigma,\Phi_g} \mathbb{E}_{\Gamma^t\sim p(\Gamma^t)}[-\log p(\Gamma)] + \beta I(X;Z) \quad (1)$$

  - Φ indicates weights and biases; Γ^t and Γ are true and predicted gender labels
  - $I(X;Z)$ is the **mutual information** between input set X and encoding set Z
  - β is a budget scaling factor for controlling information minimization
- $I(X;Z)$ is computationally challenging, find analytical upper bound $I_{max}(X;Z) \geq I(X;Z)$:

$$I(X;Z) = \int p(x,z)\log p(z|x)dxdz - \int p(z)\log p(z)dz \quad (2)$$

  - construct encoding variable $z = \sigma(c(x))\cdot\epsilon + \mu(c(x))$, where $\epsilon \sim \mathcal{N}(0,I)$
  - now $p(z|x)$ follows a Gaussian distribution $\mathcal{N}(\mu(c(x)),\sigma(c(x)))$
  - backpropagation can be efficiently performed by updating Φ_μ and Φ_σ [3]

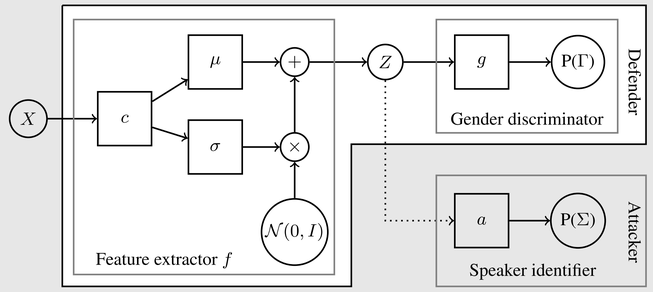


FIGURE 1: Flow chart of privacy-aware feature extraction for gender discrimination vs. speaker identification. Here f is composed of a CNN structure c and dense layers μ and σ which use stochastic sampling to transform LMBE feature set X into set Z. The MLP structure g then estimates the gender class labels' probabilities $P(\Gamma)$. The MLP structure a intercepts Z and estimates the speaker labels' probabilities $P(\Sigma)$.

- introduce variational distribution $q(z) \sim \mathcal{N}(0,I)$ and benefit from $KL(p(z)||q(z)) \geq 0$
- combine above inequality with (2) and get $I(X;Z) \leq KL(p(z|x)||q(z)) = I_{max}(X;Z)$
- Rewrite (1) as:

$$\min_{\Phi_c,\Phi_\mu,\Phi_\sigma,\Phi_g} \mathbb{E}_{\Gamma^t\sim p(\Gamma^t)}[-\log p(\Gamma)] + \beta I_{max}(X;Z) \quad (3)$$

Train attacker

- White-box attack: concatenate already trained feature extractor f with speaker identifier a
- Keep $\Phi_c, \Phi_\mu, \Phi_\sigma$ fixed and only update Φ_a
- Minimize cross-entropy between speaker labels' true $P(\Sigma^t)$ and estimated $P(\Sigma)$ probability distributions:

$$\min_{\Phi_a} \mathbb{E}_{\Sigma^t\sim p(\Sigma^t)}[-\log p(\Sigma)] \quad (4)$$

Network configuration

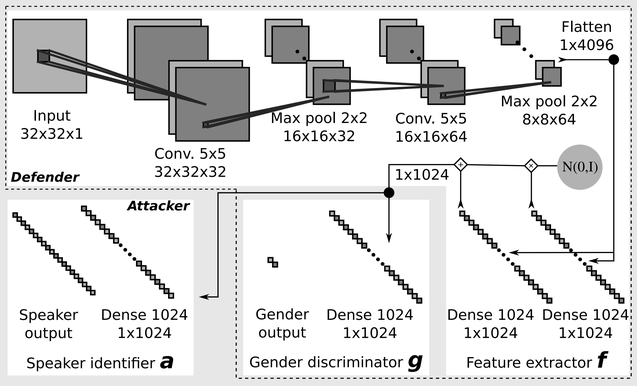


FIGURE 2: Network architecture for privacy-aware variational information feature extraction.

Experimental Results

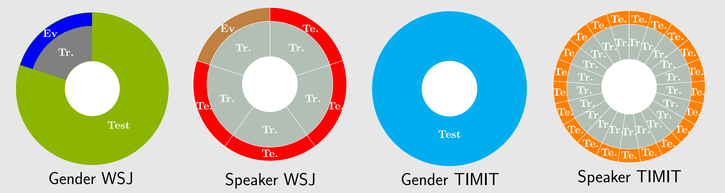


FIGURE 3: Division of training (Tr.), evaluation (Ev.) and testing (Te.) data using the WSJ corpus with 5 groups of 20 speakers and the TIMIT corpus with 21 groups of 20 speakers.

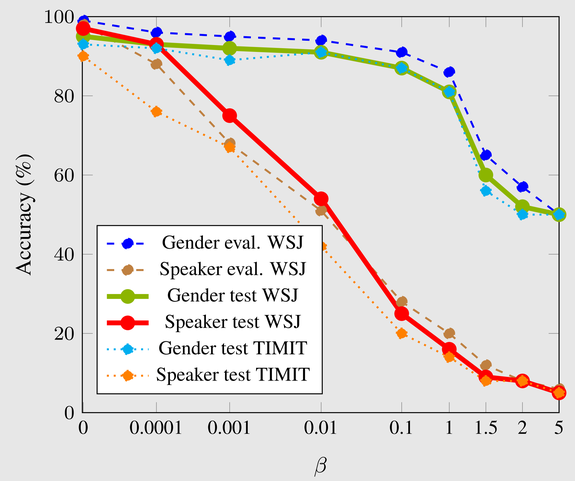


FIGURE 4: The influence of the budget scaling factor β on gender discrimination and speaker identification accuracy using the WSJ and TIMIT data sets. For $\beta = 0$ no information minimization is applied.

Conclusions and Outlook

- Speaker identification **risks** can be drastically **reduced without** significantly **deteriorating** gender discrimination **accuracy**
- Each **input** X gets **mapped to a distribution** rather than a unique Z **which** in turn, controlled by β, **ignores** as many **details** of X as possible
- Proposed **concept** can be further **expanded to** other **utility vs. privacy** applications

References

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

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

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

Research Unit FOR2457 "Acoustic Sensor Networks", work supported by DFG under grant MA 1769/6-1