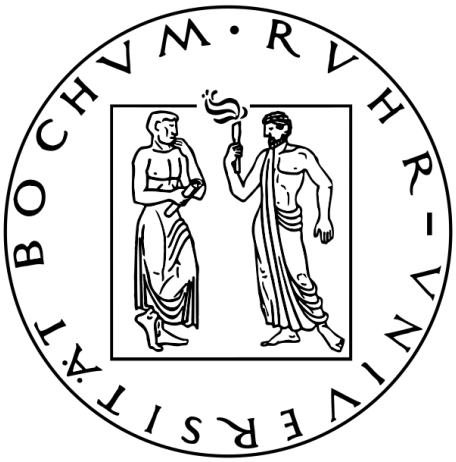
Thesis Title

Theis Name
Institute of Communication Acoustics

Ruhr-Universität Bochum



Author:	**Ravi Shah** 108018XXXXXX
Time Duration:	21.10.2023 - 01.12.2022
Submitted to:	Prof. Dr.-Ing Martin M.Sc. Lucas Bilal

Eidesstattliche Erklärung

Ich erkläre, dass ich keine Arbeit in gleicher oder ähnlicher Fassung bereits für eine andere Prüfung an der Ruhr-Universität Bochum oder einer anderen Hochschule eingereicht habe.

Ich versichere, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Die Stellen, die anderen Quellen dem Wortlaut oder dem Sinn nach entnommen sind, habe ich unter Angabe der Quellen kenntlich gemacht. Dies gilt sinngemäß auch für verwendete Zeichnungen, Skizzen, bildliche Darstellungen und dergleichen.

Ich erkläre mich damit einverstanden, dass die digitale Version dieser Arbeit zwecks Plagiatsprüfung verwendet wird.

Official Declaration

Hereby I declare that I have not submitted this thesis in this or similar form to any other examination at the Ruhr-Universität Bochum or any other institution or university.

I officially ensure that this paper has been written solely on my own. I herewith officially ensure that I have not used any other sources but those stated by me. Any and every parts of the text which constitute quotes in original wording or in its essence have been explicitly referred by me by using official marking and proper quotation. This is also valid for used drafts, pictures and similar formats.

I agree that the digital version will be used to subject the paper to plagiarism examination.

Not this English translation but only the official version in German is legally binding.

_____ _____

Datum / Date Unterschrift / Signature

Contents

List of Figures

List of Abbreviations

1 Introduction

Over the past decade, the proliferation of smart home and Internet of Things (IoT) devices has considerably enhanced the quality of our daily lives. However, these technological advancements come at a hidden cost. With the advancements in complex machine learning algorithms, there have been growing concerns about the sensitive data collected by big tech companies and the potential misuse of this information.
Speech signals encapsulate a wealth of information and can reveal a multitude of characteristics such as accent, pitch, speed, gender, health status, and the emotional state of the speaker. By analyzing the content of speech like tone, bass, timbre, articulation, rhythm, stress patterns, and phonetic nuances over time, one can discern not only the overt message being conveyed but also subtle cues that hint at the speaker's background, upbringing, and current state of mind. These cues, when interpreted correctly, can provide insights into shopping preferences, linguistic habits, social interactions, and behavioral patterns. Using advanced machine learning methodologies, a comprehensive profile of an individual user can be derived, particularly when distinct speaker identification is achievable. In reaction to growing concerns, a myriad of legal frameworks and standards have been instituted globally. Prominent among these are the European Union's General Data Protection Regulation (GDPR) [**?**], the California Consumer Privacy Act (CCPA) [**?**], and Brazil's General Personal Data Protection Law (LGPD) [**?**]. These regulations mandate corporations to integrate *privacy-by-design* principles during the development of signal processing systems. Consequently, these have amplified the urgency and significance of research in the arena of privacy-preserving signal processing.

Utility represents the enhanced capabilities and service quality offered by modern speech command services provided by corporations. However, the pursuit of this *utility* inherently involves a trade-off: the risk of data exposure, thereby compromising *privacy*. *Utility* emphasizes the functionality and efficiency of a service, aiming to maximize user benefit and experience, while *privacy* prioritizes the protection of personal information to ensure that users' data remains confidential and secure. While *utility* seeks optimization through data analysis, *privacy* cautions against potential overreach and misuse. Balancing *utility* and *privacy* is a challenge, as it aims to maximize service efficiency while safeguarding user data.

The privacy enhancement approach in this project incorporates the use of a feature extraction model, transmitting only the essential features from the raw data to the server, thereby minimizing mutual information. This method is inspired by the paper [**?**], aiming to optimize accuracy in gender discrimination (*utility*) while ensuring the results for speaker identification remain non-distinctive to maintain *privacy*. This approach draws inspiration from the concept of the *Variational Information Bottleneck* (VIB) [**?**], which is motivated

by the *Information Bottleneck* (IB) principle [**?**]. This principle aims to retain only the most relevant information for a task, thereby reducing the risk of exposing unnecessary data. The VIB method incorporates elements of the *variational autoencoder*, a type of neural network(NN) that can generate new, similar data based on the training set. This technique encodes compact data representations and employs a *re-parametrization* method [**?**], introducing variability and allowing for stochastic sampling during backpropagation. This helps optimize the balance between *utility* and *privacy* by ensuring that the model does not overfit to sensitive information.

In this project, we deploy a deep neural network(DNN)-based, privacy-aware feature extraction technique for gender recognition(GR) and speaker identification(SI), as detailed in the referenced paper [**?**]. The primary goal is to balance the trade-off between GR and SI, thereby mitigating potential data exposure risks by minimizing the amount of data transferred over the channel. Our investigation into this scenario aims to enhance the classifier's GR capabilities while simultaneously restricting access to precise speaker identity information.

The aim of this work is to examine the effectiveness of the trainable acoustic frontend, Per Channel Energy Normalization (PCEN), in GR vs SI tasks. To evaluate this, we compared the performance of Per Channel Energy Normalization(PCEN) with that of Log Mel Band Energy (LMBE) as a feature extractor. We evaluated the performance of PCEN with both trainable and non-trainable parameters. We utilized the *Librispeech* dataset for our experiment and extracted high-level features by training the GR trust model. Next, applying the VIB principle, we established a probabilistic mapping between the input data and a compact latent space through the *re-parametrization trick*, as discussed in **??**. By leveraging the VIB principle and gradually increasing the budget scaling factor β, our goal is to identify a β at which the features compress in such a way that GR remains achievable, but SI becomes challenging. Furthermore, we investigated the accuracy of the SI by training the high-level feature extractor, referred to as the Trust model, using the *Sonyc UST* dataset.

Before transmitting information to the server side, a scheme can be applied at the node level that employs deep neural network (DNN)-based feature extraction. We utilize a privacy-centric variational information feature extraction technique that deliberately produces lossy results, anchoring on Mutual Information (MI) minimization. Our analysis focuses on its impact on the delicate balance between GR accuracy (*utility*) and SI (*privacy*). Striking a balance between these two elements is crucial; optimizing one often compromises the other. We examine the loss function and the budget scaling factor β, which influences the trade-off between *utility* and *privacy*, for our proposed model. Experimental results have demonstrated that our method substantially mitigates the associated privacy risks without causing significant degradation in *utility*. Impressively, these outcomes are also consistent with speech sources not previously observed. In the context of wireless acoustic sensor networks (WASNs), implementing node-level feature extraction is especially beneficial as it offers the potential to decrease bandwidth consumption, distribute computational load across the network, and hence enhance *privacy* by incorporating aggregation techniques.

The structure of the thesis is organized as follows:

- **Chapter 2: Feature Extraction** provides an overview of various techniques, emphasizing the detailed discussion on Log Mel Band Energy (LMBE) and Per Channel Energy Normalization (PCEN).
- **Chapter 3: Fundamentals of Deep Learning Networks** delves into the core concepts of DNN, covering topics such as activation functions, optimization methods, Convolutional Neural Networks (CNN), and the training processes of DNNs.
- **Chapter 4: Privacy-preserving Feature Extraction** explores methodologies and principles for preserving privacy in feature extraction, including discussions on the Variational Autoencoders (VAE), Variational Information Bottleneck(VIB) principle and the reparametrization trick.
- **Chapter 5: Experimental Setup and Results** details the experimental framework for both the *Librispeech* and *SONYC UST* datasets and presents the findings obtained from the experiments.

Conclusions

Future Work

- **Exploration of PCEN Parameters:** A promising avenue for future research lies in the exploration of specific parameters within PCEN. Fine-tuning these parameters could potentially enhance the performance difference between the GR and SI tasks, ensuring optimal privacy preservation.
- **Modifying PCEN for higher dynamic range:** One potential direction for enhancement is to modify PCEN to achieve a higher dynamic range. This could address the observed limitations of PCEN in terms of information loss and concurrent performance degradation.
- **Evaluation on Diverse Datasets:** To ascertain the generalizability of our findings, it would be beneficial to test them on a variety of datasets. This approach would offer a more comprehensive understanding of the strengths and limitations of both PCEN and LMBE in different contexts.
- **Exploring MobileNetV2 for SONYC UST Dataset**: In future endeavors, it would be worthwhile to investigate the application of the *MobileNetV2* architecture for the SONYC UST dataset. Given MobileNetV2's efficiency in terms of computational resources and its adaptability to various tasks, it presents a promising avenue for enhancing classification performance. The compact nature of MobileNetV2 might offer a balance between accuracy and computational efficiency, making it particularly suitable for real-time or on-device applications related to the SONYC UST dataset.