Title: Enhancing Speaker Anonymization Using Disentanglement Learning

Abstract:

Speaker Anonymization is the process of making speech data unlinkable to individual identities by concealing voice characteristics, accents, and speaking styles. This technique addresses privacy concerns from smart home voice assistants that collect sensitive speech data. To standardize the evaluation process for speaker anonymization, the Voice Privacy Challenge 2020 (VPC2020) was introduced in 2020, with VPC2024 presenting six baseline systems (B1–B6). These six baseline systems consist of 5 different Voice Conversion (VC) and 1 Digital Signal Processing (DSP) approaches.

In the first part of our work, we examine VPC2024 and focus on B3 and B5 VC baseline systems. The B3 utilizes phonetic transcriptions and a GAN to generate pseudo-speaker embeddings, modifying pitch, energy, and prosodic features before synthesizing anonymized speech with FastSpeech2 and HiFi-GAN. The B5 approach uses a wav2vec2-based ASR model with vector quantization to extract linguistic content and one-hot vectors to represent speaker embeddings, which HiFi-GAN then employs to synthesize anonymized speech. To enhance these systems, we employ Disentanglement Learning to separate linguistic and paralinguistic features for acoustic anonymization. For B3, we integrated emotion embeddings and speaker embedders like WavLM and ECAPA2, and applied prosody manipulation techniques, resulting in a 15.5\% improvement in Unweighted Average Recall (UAR) on the IEMOCAP dataset. For B5, we implemented the Mean Reversion method and additive white Gaussian noise (AWGN) on prosody, enhancing privacy with a 32.2\% improvement in Equal Error Rate (EER) on the Librispeech dataset and securing third place in privacy preservation at VPC2024. Additionally, we adapted the disentanglement-based NaturalSpeech3 FACodec for speaker anonymization, demonstrating promising results compared to β-VAE.

Next, we explore the Household Speaker Recognition application to support our future work of Speaker Anonymization in smart homes. In this work, we develop a probabilistic back-end for online speaker recognition and clustering, achieving a 4\% improvement in EER on VoxCeleb1 and an 8\% relative reduction in Diarization Error Rate (DER) on the AMI dataset { compared to previous approaches based on cosine scoring}. This framework will allow us to apply disentanglement-based Speaker Anonymization using enhanced B3 and B5 to improve privacy protection.

Chapters & Questions

1. Introduction

1.1. Background

- What is Speaker Anonymization?
- What is Disentanglement?
- How are Disentanglement Learning and Anonymization connected?
- What is the common application scenario of Speaker Anonymization -> Smart Speaker -> Household speaker recognition

1.2. Motivation

- Why is Disentanglement Learning used for anonymization?
- How Disentanglement Learning helps to achieve better anonymization?

1.3. Contribution

- How are we solving it?

1.3.1.1. Disentanglement Learning for Speaker Anonymization:

- Emotion Embedding to preserve emotions
- MeanReversion + AWGN for F0 to enhance privacy protection
- NaturalSpeech3 FACodec for Speaker Anonymization with emotion preservation

1.3.1.2. Household Speaker Recognition Scenario:

- What is proposed? (briefly)

1.4. Outline of the Thesis

2. Literature Review

2.1. Speaker Anonymization

2.1.1. General information

- Why did this direction become important?
- How does a general pipeline look like?
- What kind of Attacker/Threat models exist?
- What kind of Evaluation metrics are used for anonymization?
- Which Datasets are used?
- How do current methods approach Speaker Anonymization?

2.1.2. VoicePrivacy Initiative

- Why was the VoicePrivacy Initiative started?
- What is VoicePrivacy Challenge (VPC)?
- What are the differences between VPC2020, VPC2022, VPC2024?
- How have privacy requirements been evolving?

2.1.3. Attacker Models

- What kind of Attacker/Threat models exist?
- How to define a successful attack?
- How to define a failed attack?

2.1.4. Evaluation metrics

- What kind of Evaluation metrics are used for anonymization?
- How do these metrics correspond to each other? Explain Utility vs Privacy trade-off

2.1.5. Datasets and Corpora

- Which datasets are used?
- Why these datasets?

2.1.6. VPC2024 Baseline Models

- How do current methods approach Speaker Anonymization?

2.2. Connection of Speaker Anonymization Approaches to Voice Conversion and TTS
- What are the differences/similarities between Speaker Anonymization and Voice Conversion

2.3. Summary of the chapter

3. Disentanglement-based Approaches for Anonymization

3.1. Disentanglement Learning in Speech

3.1.1. Overview of Disentanglement Learning in Speech
- How does the pipeline look like?
- Whats the main idea behind Disentanglement Learning?

3.1.2. Methods and Approaches for Disentanglement in Speech
- How do researchers approach Disentanglement Learning in Speech?

3.1.3. Applications of Disentangled Speech Representations
- What are applications of Disentangled Speech Representations?

3.1.4. Two recent approaches for Disentanglement Learning

3.1.5. Challenges in Disentanglement Learning for Speech
- Why is Disentanglement Learning in Speech Challenging?

3.2. Disentanglement approaches for VPC2024

3.2.1. Baseline Systems
- What is the related work?
- What are the existing methods?
- Why do those methods need to be improved?

3.2.2. Our Methods
- Modifications of B3
- Disentanglement-based models
- Mean Reversion F0 for the B5 system

3.2.3. Experiments
- What are the corpora and evaluation metrics?
- What are our experimental setups?
- What are our insights?

3.3. Conclusion

4. Household Speaker Recognition Scenario

4.1. Introduction
- What is the motivation?
- How can current approaches be improved?
- What is proposed in this work?

4.2. Background
- What is PLDA?
- How is PLDA connected to Multi-enrollment speaker recognition?
- What is PSDA?
- What is Multi-enrollment verification?
- How is Multi-enrollment verification performed?

4.3. Online probabilistic speaker clustering
- What is the general definition of Online Clustering?
- How is it connected to Multi-Enrollment speaker recognition?

- What is the idea under the proposed algorithm?

4.4. Experiments
- How is Multi-enrollment verification performed
- Which corpus and metrics are used to evaluate performance?
- How is the model trained?
- What are insights?

4.5. Conclusion

5. Conclusion and Future Work
   5.1. Conclusion
   5.2. Room for Improvement
   5.3. Future work