NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Enhancing Speaker Anonymization Using Disentanglement Learning

Qualifying Examination Report

Submitted to the College of Computing and Data Science

of the Nanyang Technological University

by

Nikita Kuzmin

Supervisors: Prof. Chng Eng Siong and Dr. Hardik B. Sailor

Jan, 2025

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

Jan. 2025

Date

Nikita Kuzmin

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

......Jan. 2025......

Date

Prof Chng Eng Siong

Authorship Attribution Statement

This thesis contains material from 2 papers in which I am listed as an author.

Chapter 3 is published as **Kuzmin, N.**, Luong, H.-T., Yao, J., Xie, L., Lee, K.A., Chng, E.-S. (2024) NTU-NPU System for Voice Privacy 2024 Challenge. Proc. 4th Symposium on Security and Privacy in Speech Communication, 72-79, doi: 10.21437/SPSC.2024-13.

- I co-designed the experiments with Dr. Luong and Jixun Yao.
- We prepared the codebase with Dr. Luong.
- We prepared the manuscript drafts with Dr. Luong. The paper was revised by Prof. Chng, Prof. Lee and Prof. Lei Xie.

Chapter 4 is published as A. Sholokhov*, **N. Kuzmin***, K. A. Lee and E. S. Chng, ”Probabilistic Back-ends for Online Speaker Recognition and Clustering,” ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10097032.

The contributions of the co-authors are as follows:

- I co-designed the experiments with Dr. Sholokhov and prepared the codebase.
- We prepared the manuscript drafts with Dr. Sholokhov. The paper was revised by Prof. Chng and Prof. Lee.

Jan. 2025
....
Date

....
Nikita Kuzmin

List of Figures

List of Tables

Contents

List of Abbreviations

Automatic Speaker Verification	ASV
Automatic Speech Recognition	ASR
Convolutional Neural Network	CNN
Deep Neural Network	DNN
Equal Error Rate	EER
Voice Activity Detection	VAD
Emotion Recognition	ER
Mel-frequency Cepstral Coefficients	MFCCs
Generative Adversarial Network	GAN
Gaussian Mixture Model	GMM
Universal Background Model	UBM
Joint Factor Analysis	JFA
Time-Delay Neural Network	TDNN
Recurrent Neural Network	RNN
Long Short-Term Memory	LSTM
Fully Connected	FC
Probabilistic Linear Discriminant Analysis	PLDA
Linear Discriminant Analysis	LDA
General Data Protection Regulation	GDPR
False Rejection	FR
False Acceptance	FA
False Reject Rate	FRR
False Acceptance Rate	FAR
Speech Emotion Recognition	SER
Self-Supervised Learning	SSL
Detection Cost Function	DCF
Text-to-Speech	TTS
Variational Autoencoder	VAE
Generative Adversarial Network	GAN
Voice Conversion	VC
VoicePrivacy Challenge	VPC
Natural Language Processing	NLP
Gradient Reversal Layer	GRL
Augmentation	AUG

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.

Chapter 1

Introduction

1.1 Background

Advancements in automatic speech recognition (ASR) and text-to-speech (TTS) technologies have started the development of voice assistants like Amazon's Alexa, Apple's Siri, and Google Assistant [1]. These devices offer convenience and new modes of interaction in homes and workplaces. However, they also introduce substantial privacy and security concerns, as the continuous collection and centralized storage of user speech data can expose sensitive personal information [2, 3, 4]. High-profile incidents of unauthorized data access since 2017 have heightened public awareness and underscored the urgent need for robust privacy protections and stricter regulations, such as the General Data Protection Regulation (GDPR) in the European Union [5]. In environments like smart homes, multi-enrollment speaker recognition systems are commonly employed to manage interactions with multiple users. These systems create robust voice profiles by comparing multiple enrollment utterances with multiple test utterances, enabling accurate identification and differentiation between users. While this functionality enhances personalized experiences, it simultaneously amplifies privacy risks by increasing the potential for misuse or unauthorized access to sensitive speech data [6]. Therefore, integrating effective anonymization techniques within multi-enrollment frameworks is essential to mitigate these risks without compromising user convenience.

Addressing these privacy challenges requiring anonymization methods that integrate seamlessly with existing data processing systems without demanding extensive modifications. Unlike other privacy-preserving methods such as encryption or federated learning, which may require significant changes to data pipelines and complicate the creation of large, usable speech datasets [7, 8, 9, 10], speaker anonymization transforms speech into a privacy-preserving format that aligns with current pipelines. This practicality is essential for widespread adoption, ensuring that privacy protections do not affect the deployment and effectiveness of speech technologies. Figure 1.1 illustrates a household scenario where smart devices are interacting with cloud services. An attacker attempts to intercept and extract speaker identities (Speaker ID) from the speech data transmitted to or stored in the cloud. The Speaker Anonymization system, embedded within the smart device, prevents the attacker from gaining access to sensitive speaker information while still allowing the speech data to be useful for utility tasks like ASR or Speech Emotion Recognition (SER). This approach addresses security concerns while maintaining the functionality of speech-based services.

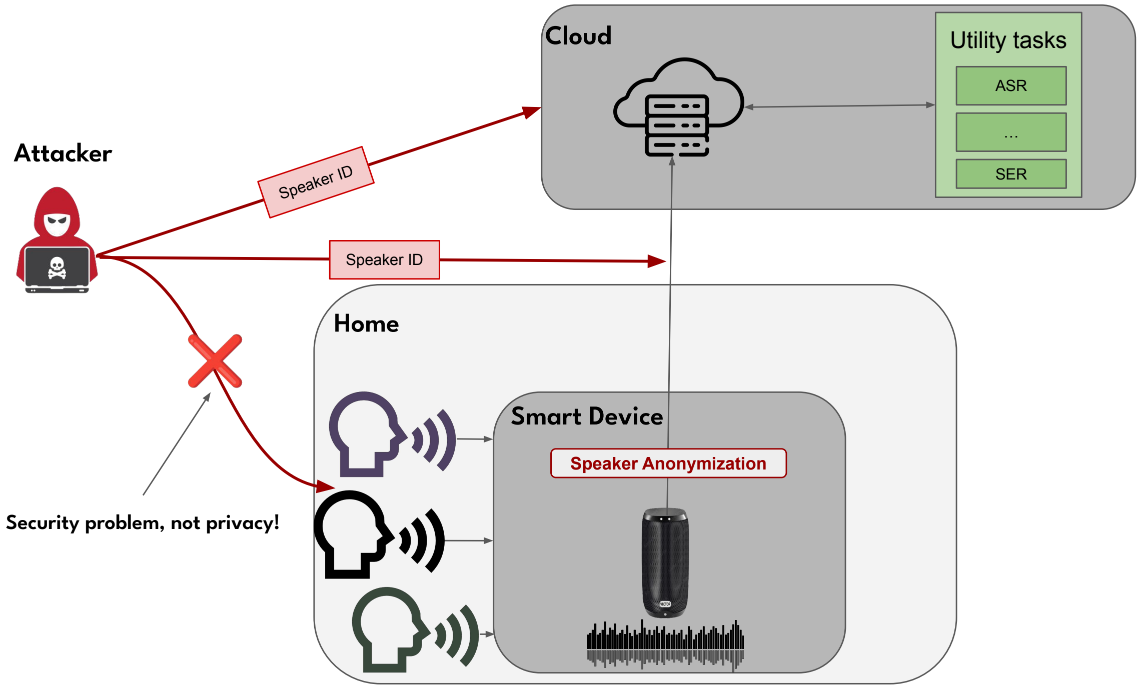


Figure 1.1: Speaker Anonymization in a household scenario.

Protecting user privacy in speech-enabled technologies has become increasingly crucial as verbal communication systems become important in everyday life. Speech signals inherently contain sensitive personal information, including speaker identity, emotional states, age, and gender, which can be exploited to compromise individual privacy [2, 3, 4]. In response to these growing privacy concerns, speaker anonymization has emerged as a pivotal area of research. Speaker anonymization techniques aim to protect user privacy by removing identifiable information from speech signals, effectively concealing speaker-specific features while preserving the essential linguistic content necessary for communication [11]. A promising strategy within speaker anonymization is disentanglement learning, which seeks to separate the complex factors of variation in speech data into independent components [12, 13]. By disentangling linguistic representations from paralinguistic ones—such as speaker timbre, emotion, age, and gender—disentanglement learning enables precise manipulation of speech signals to anonymize specific aspects without degrading the overall utility of the speech [14, 15, 16]. This approach not only enhances privacy protection but also maintains the functionality and quality of speech-enabled services, striking a balance between user privacy and system performance.

Ongoing initiatives like the VoicePrivacy Challenge [17] highlight the efforts to enhance speaker anonymization through advanced voice conversion systems, which aim to reduce linkability and maintain linguistic and emotional integrity [18]. However, developing robust anonymization methods that can withstand sophisticated attacks while preserving speech utility remains a significant challenge [19]. Continued research in disentanglement learning and speaker anonymization is crucial for advancing privacy-preserving technologies in the ever-evolving landscape of speech-enabled applications.

1.2 Motivation

As speech-enabled technologies become essential to daily life, protecting user privacy has become paramount. Speaker anonymization plays a crucial role in this context by removing identifiable information from speech signals, safeguarding sensitive attributes such as speaker identity, emotions, age, and gender [16]. This is especially important for devices like smart speakers that continuously collect and store users' speech data, heightening the risk of unauthorized access and misuse.

A promising method for speaker anonymization is disentanglement learning, which separates the complex variations in speech data into distinct components [12, 13]. By isolating linguistic content from paralinguistic features, disentanglement learning allows for the precise manipulation of specific speech attributes. This enables the effective anonymization of speaker-specific information while preserving the intelligibility and naturalness of the speech, ensuring that the utility of speech-enabled services remains intact [20].

Moreover, disentanglement learning offers the flexibility to choose which speaker-specific features to conceal based on specific privacy requirements. This targeted approach ensures that only the necessary attributes are anonymized, maintaining the balance between privacy protection and speech utility. However, disentanglement learning approaches remain underexplored for speaker anonymization, particularly in the disentanglement of emotions, presenting a significant opportunity for further research [21, 22].

In conclusion, the combination of disentanglement learning and speaker anonymization provides a robust framework for enhancing privacy in speech-enabled technologies. By allowing selective concealment of speaker-specific features and addressing the gaps in current research, particularly in emotion disentanglement, this approach holds great potential for developing effective and adaptable privacy-preserving solutions in the evolving landscape of speech technologies.

1.3 Contribution

In the first part of our study [23], we concentrate on speaker anonymization by employing disentanglement learning approaches to effectively separate linguistic and paralinguistic features, thereby anonymizing acoustic information. Specifically, we focus on enhancing systems B3 and B5 from the VoicePrivacy Challenge 2024 [24], which involve the disentanglement of prosody, speaker identity, and content. For system B3, we integrated emotion embeddings and advanced speaker embedders such as WavLM [25] and ECAPA2 [26], and explored various prosody manipulation techniques to improve anonymization while retaining emotional and content-related features. In the case of system B5, we applied the Mean Reversion method and Additive White Gaussian Noise (AWGN) to prosody, enhancing privacy without compromising the utility of the speech data.

These modifications led to significant performance improvements: the enhanced B3 system achieved an average increase of 15.5% in Unweighted Average Recall (UAR) for emotion recognition on the IEMOCAP [27] dataset across both development and test splits. Furthermore, our modified B5 system demonstrated a 32.2% improvement in Equal Error Rate (EER) on the LibriSpeech [28] development and test splits, securing third place in privacy preservation at VPC2024. Additionally, we adapted the disentanglement-based NaturalSpeech3 FACodec [29] for the speaker anonymization task and compared its performance to β-VAE [30], with FACodec showing promising results for voice privacy. These enhancements collectively advance the state-of-the-art in speaker anonymization by effectively balancing the preservation of essential speech attributes with the protection of personal identity information.

In the second part of our study [31], we address the challenges associated with multi-enrollment speaker recognition in online scenarios. While our primary emphasis is on anonymization, we briefly explore efficient aggregation methods for handling an increasing number of speech segments per speaker. This is particularly relevant for applications such as household smart devices and online speaker diarization, where managing multiple enrollments is essential. By evaluating existing cosine scoring methods and introducing a constrained Probabilistic Linear Discriminant Analysis (PLDA) [32] with spherical covariance matrices, we offer improved performance and computational efficiency. Additionally, our probabilistic back-end, based on incremental variational Bayesian inference, provides soft, probabilistic decisions for incoming speech segments, enhancing the system's adaptability to varying uncertainties. This approach resulted in a 4% improvement in EER on the custom evaluation protocols from VoxCeleb1 dataset [33] and an 8% relative reduction in Diarization Error Rate (DER) on the AMI dataset [34]. This aspect is preliminary work to our anonymization focus, we will next extend this work by applying disentanglement-based anonymization techniques.

1.4 Outline of the Thesis

The remaining part of this work is structured as follows:

- Chapter 2 consists of literature review. To be more specific, it introduces speaker anonymization and Voice Privacy Challenge initiative.
- Chapter 3 reviews disentanglement learning approaches for anonymization task. Furthermore, proposes modifications of several disentanglement-based anonymization systems improving privacy and utility metrics with focus on emotion recognition.
- In Chapter 4, we evaluate scoring back-ends for multi-enrollment verification using large-margin embeddings. Additionally, we propose an alternative to cosine scoring and a probabilistic back-end for online speaker recognition and clustering.
- Chapter 5 provides conclusion and reveals possible directions for future work.

Chapter 2

Literature Review

In this chapter, we first introduce an overview of speaker anonymization, highlighting the privacy implications of voice-based technologies and the motivations for anonymizing speaker identity in Section 2.1. In detail, the chapter explores the structure and evaluation of anonymization systems, including utility and privacy metrics, and introduces the VoicePrivacy Challenge (VPC) [11, 17, 24] framework as a standardized benchmark for evaluating anonymization methods. Additionally, a comprehensive literature survey of current speaker anonymization techniques is presented, focusing on methods used across different VPC editions, such as digital speech processing [35], x-vectors [36, 37], neural codecs [38], and generative models [37, 39, 40, 41]. The evaluation frameworks and metrics, including EER and WER, are discussed to illustrate the trade-off between privacy and utility. Overall, this chapter provides foundational knowledge of speaker anonymization and disentanglement, laying the groundwork for the contributions in Chapter 3 and Chapter 4.

2.1 Speaker Anonymization

2.1.1 General Information

The rapid advancement of speech-based technologies has transformed human-computer interaction, voice biometrics, and natural language processing (NLP). These innovations have enabled more intuitive and efficient interfaces, improving user experiences across various applications. However, the growing reliance on voice data has raised significant privacy concerns. Voice recordings can reveal sensitive information such as a speaker's identity, emotional state, and demographic details like age, gender, and ethnicity, making them vulnerable to unauthorized access.

This concern has gained attention due to the increasing collection of speech data and the associated risks of revealing personal information. The European General Data Protection Regulation (GDPR) [42] further highlights the need for robust privacy protections, as it mandates strict measures for safeguarding personal data, including biometric information from voice recordings. Speaker anonymization has thus emerged as a crucial field aimed at protecting individuals' privacy while maintaining the functionality of voice-based systems.

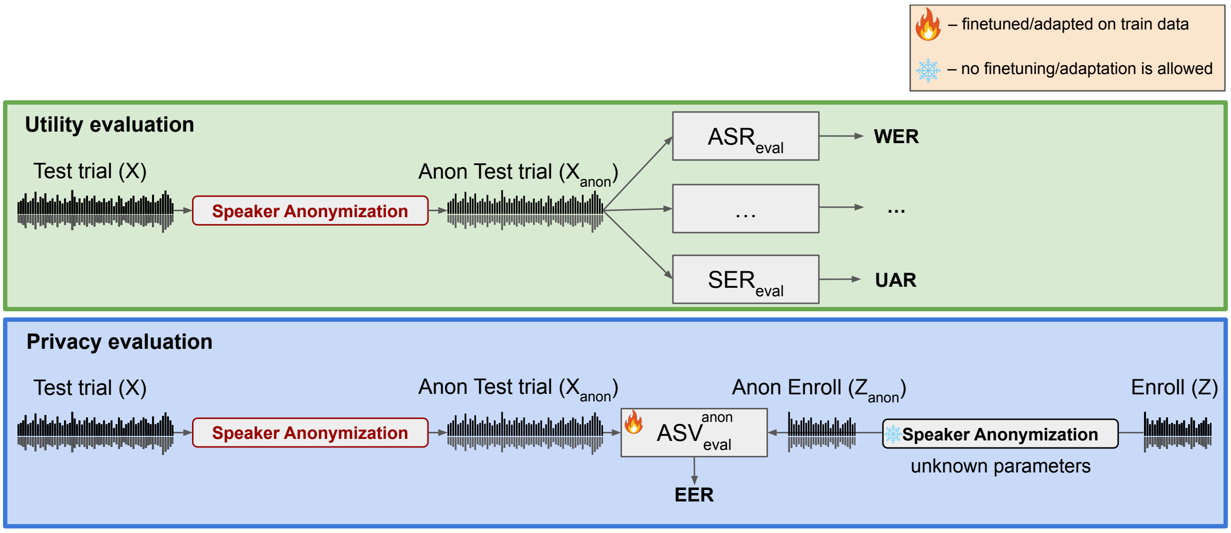


Figure 2.1: Speaker Anonymization Evaluation pipeline scheme with Semi-Informed attacker.

Figure 2.1 presents a general evaluation pipeline for speaker anonymization, which is divided into two components: utility and privacy. Privacy evaluation is crucial to ensure that speaker anonymization effectively prevents the identification of the original speaker, safeguarding personal data and confidentiality. Utility evaluation is essential to confirm that the anonymized speech remains intelligible and retains linguistic and prosodic information for practical applications such as ASR, SER, etc. During the utility evaluation, the speaker's test trial (X) is processed through the anonymization system, resulting in anonymized speech (X_{anon}). This anonymized speech is then input into utility models such as ASR evaluation (ASR_{eval}) and SER evaluation (SER_{eval}), which are pre-trained models for the respective tasks on the unanonymized data. For privacy evaluation, the attacker model ($\text{ASV}^{\text{anon}}_{\text{eval}}$) is introduced, leveraging both the anonymized enrollment data (Z_{anon}) and the anonymized test trial (X_{anon}) to attempt to assess if Test trial (X) and Enroll (Z) are from the same or different speakers. In this semi-informed attack scenario, the attacker model is aware of the anonymization method and has been fine-tuned on anonymized data. The attack is deemed successful if the model correctly identifies that the anonymized test trial and anonymized enrollment data originate from the same speaker, thereby breaching privacy.

As examples, in VPC2024, the ASR_{eval} model is based on *wav2vec2-large-960h-lv60-self* [43] and is finetuned on *LibriSpeech-train-960* [44] using the SpeechBrain recipe [45]. The SER_{eval} model utilizes *wav2vec2* and is trained on specific folds of the IEMOCAP dataset [27]. For the $\text{ASV}_{\text{eval}}^{\text{anon}}$ model, ECAPA-TDNN [46] is employed and trained on LibriSpeech protocols. More details on attacker models can be found in Subsection 2.1.3.

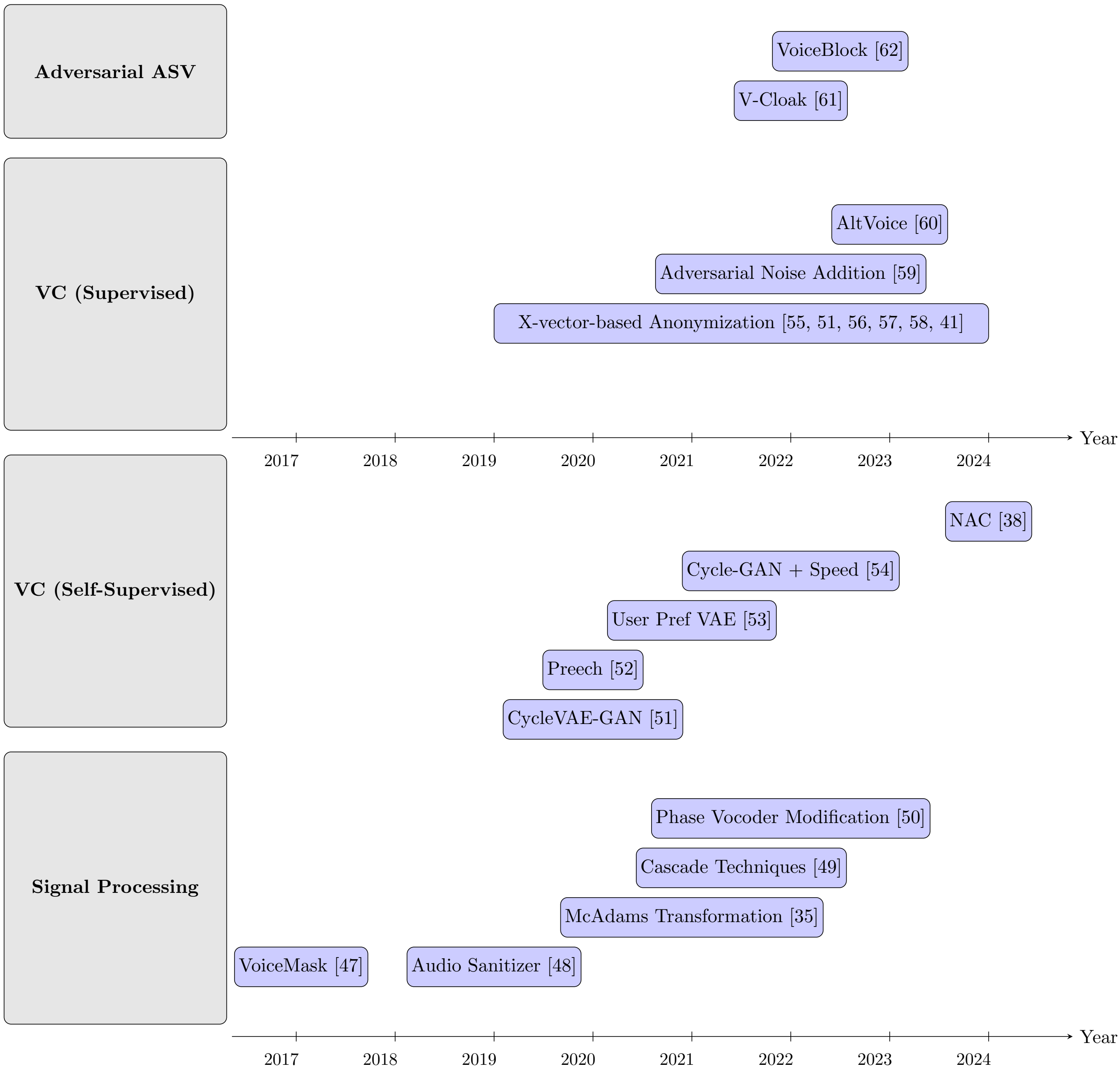


Figure 2.2: Timeline of Speaker Anonymization approaches.

Next, in this overview we briefly discuss various aspects of attacker models, as illustrated in Figure 2.1, which presents a classification of different attacker models based on their knowledge and target features. Attacker models vary depending on their level of knowledge about the defender's system (anonymization system) [17, 63, 24]. These can range from attackers with zero knowledge, referred to as "ignorant attackers," to those with partial knowledge, such as ”lazy-informed” and ”semi-informed” attackers, and finally to attackers with complete knowledge, labeled as ”informed attackers.” Additionally, attackers differ by the specific features they aim to compromise. These target features include sensitive information such as Speaker Identity (Speaker ID), gender, accent, or other speaker-specific characteristics. In some cases, attackers may target multiple features simultaneously, increasing the complexity and potential risk to privacy. Understanding these variations in attacker models is crucial for evaluating the robustness of speaker anonymization systems against such multi-targeted threats.

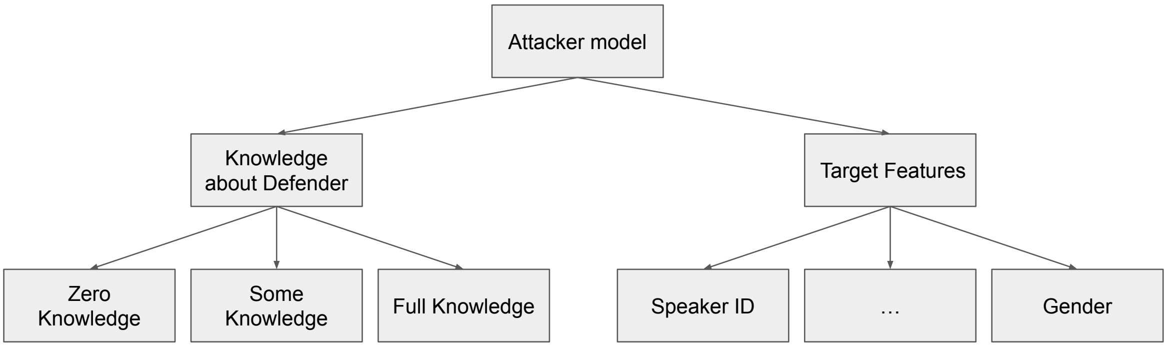


Figure 2.3: Classification of Attacker Models.

Finally, we introduce the main metrics. As was discussed before, there is a trade-off between privacy and utility. The main privacy metric is EER [24], which is calculated by the formula provided in Equation 2.1

$$EER = F_{\text{fa}}(\theta_{EER}) = F_{\text{miss}}(\theta_{EER}), \tag{2.1}$$

where $F_{\text{fa}}(\theta)$ represent the false alarm rate and $F_{\text{miss}}(\theta)$ represent the miss rate at a given threshold θ. The Equal Error Rate (EER) [24] is defined as the value of the threshold θ_{EER} where the false alarm and miss rates are equal.

As for utility, WER [24] is used. To assess the accuracy of a predicted transcript, it is compared against a reference transcript. The WER is a metric used to quantify the discrepancies between the two transcripts. This metric accounts for substitution, deletion, and insertion errors. Substitutions occur when a word is transcribed incorrectly, insertions refer to added words, and deletions are words that are omitted from the transcription. The WER is calculated using the following formula in Equation 2.2

$$WER = \frac{N_{\text{sub}} + N_{\text{del}} + N_{\text{ins}}}{N_{\text{total}}}, \tag{2.2}$$

where N_{sub}, N_{del}, N_{ins} – amount of substitutions, deletions, insertions correspondigly, N_{total} – total amount of words in the reference transcript.

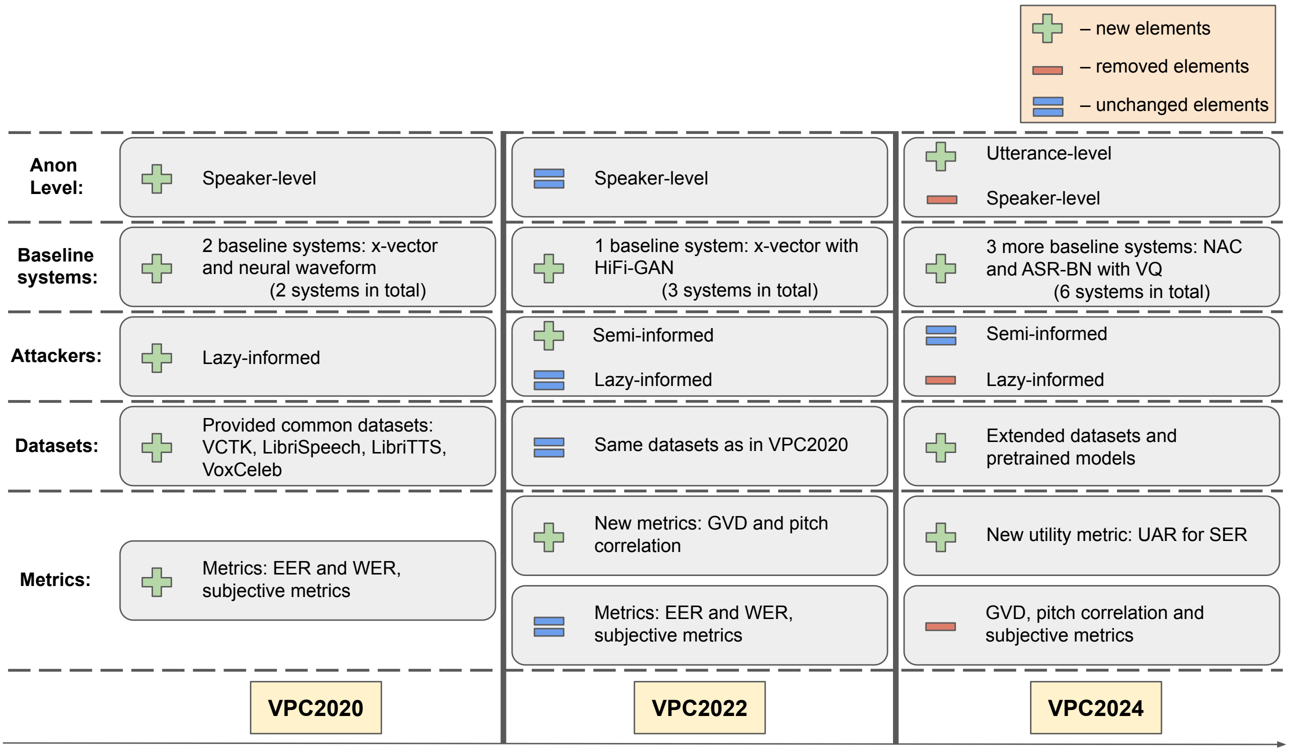
Finally, we will provide structured overview of current approaches to Speaker Anonymization in Table 2.1. Additionally, the timeline of approaches is illustrated on Figure 2.2.

Table 2.1: Overview of Speaker Anonymization Approaches.

Method	Category	Description
VoiceMask [47]	Signal Processing	Utilizes VTLN-based frequency warping to alter the spectral envelope, aiming to modify perceived speaker identity.
Audio Sanitizer [48]	Signal Processing	Modifies pitch, tempo, and pauses, adding white noise post MFCC-based waveform resynthesis.
McAdams Transformation [35]	Signal Processing	Uses McAdams coefficient for timbre modification, expanding/contracting harmonic frequencies to shift speaker identity.
Cascade Techniques [49]	Signal Processing	Cascades resampling, modulation spectrum smoothing, McAdams transformation, and chorus, with combinations achieving varying effectiveness.
Phase Vocoder Modification [50]	Signal Processing	Uses phase vocoder for time-scale compression/stretching, evaluated under multiple attacker models for privacy estimation.
CycleVAE-GAN [51]	Voice Conversion (Self-Supervised)	Extracts linguistic features via VAEs and anonymizes with GANs and a one-hot vector.
User Preference VAE [53]	Voice Conversion (Self-Supervised)	VAE that removes user-specified characteristics from speech signals.
Preech Pipeline [52]	Voice Conversion (Self-Supervised)	Ensures anonymization at multiple linguistic levels with PPGs and ASR for privacy.
Neural Audio Codec (NAC) [38]	Voice Conversion (Self-Supervised)	Applies neural audio codec language modeling with quantized codes and transformer-based token generators to re-synthesize speech, balancing privacy and utility.
Cycle-GAN with Speed Perturbation [54]	Voice Conversion (Self-Supervised)	Converts voice to opposite gender using Cycle-GAN and speed perturbation techniques.
X-vector-based Anonymization [22]	Voice Conversion (Supervised)	Alters x-vectors and uses ASR-BN features for speech synthesis; serves as a baseline in VoicePrivacy Challenge.
Adversarial Noise Addition [59]	Voice Conversion (Supervised)	Adds Laplace noise to ASR-BN and F0 features to achieve differential privacy guarantees.
AltVoice [60]	Voice Conversion (Supervised)	Converts speech to text tokens and synthesizes new speech using TTS for full unlinkability.
Random Dense Cluster Averaging [55]	X-Vector Modification	Averages random x-vectors from a dense cluster to generate target x-vectors, balancing privacy and utility under lazy-informed attacker.
One-Hot Component Modification [51]	X-Vector Modification	Randomly alters each component in one-hot speaker embeddings, increasing privacy while maintaining speaker distance.
GMM with PCA for Naturalness [56]	X-Vector Modification	Generates x-vectors via Gaussian Mixture Model with PCA, reflecting real-world x-vector distributions for naturalness.
SVD and Statistical Regression [57]	X-Vector Modification	Uses Singular Value Decomposition and regression to create target x-vectors, focusing on privacy improvements.
Adversarial Learning for Disentanglement [58]	X-Vector Modification	Enhances x-vector disentanglement from gender and accent using adversarial learning, aiming to preserve privacy utility balance.
Wasserstein GAN Target Vector Generation [41]	X-Vector Modification	Employs Wasserstein GAN to produce x-vectors mimicking natural distributions, preserving audio naturalness and privacy.
Noise in Salient Regions [64]	ASR-BN Modification	Adds noise selectively in ASR-BN regions identified by a privacy-risk estimator, improving privacy at the feature level.
V-Cloak [61]	Adversarial ASV Attack	Modulates bottleneck features of a Wave-U-Net model to adjust speaker characteristics.
VoiceBlock [62]	Adversarial ASV Attack	Applies time-varying finite impulse response filters to create imperceptible perturbations.

2.1.2 VoicePrivacy Initiative

The evaluation of anonymization approaches faced significant challenges due to the variety of attackers, metrics, datasets, and scenarios, and the lack of a standardized evaluation framework made comparisons between methods difficult. To address these privacy challenges, the VoicePrivacy Challenge (VPC) [11] was introduced to promote the development of robust voice anonymization techniques. The goal of VPC is to anonymize speaker identities while preserving the usability of speech data for tasks like automatic speech recognition (ASR) and natural language understanding. By offering standardized datasets, protocols, and evaluation metrics, VPC provides a unified platform for researchers to benchmark and improve their anonymization systems. The evolution of VPC series from 2020 to 2024 is illustrated in Figure 2.4.



✚ – new elements
🟥 – removed elements
🟦 – unchanged elements

	VPC2020	VPC2022	VPC2024
Anon Level:	✚ Speaker-level	🟦 Speaker-level	✚ Utterance-level 🟥 Speaker-level
Baseline systems:	✚ 2 baseline systems: x-vector and neural waveform (2 systems in total)	✚ 1 baseline system: x-vector with HiFi-GAN (3 systems in total)	✚ 3 more baseline systems: NAC and ASR-BN with VQ (6 systems in total)
Attackers:	✚ Lazy-informed	✚ Semi-informed 🟦 Lazy-informed	🟦 Semi-informed 🟥 Lazy-informed
Datasets:	✚ Provided common datasets: VCTK, LibriSpeech, LibriTTS, VoxCeleb	🟦 Same datasets as in VPC2020	✚ Extended datasets and pretrained models
Metrics:	✚ Metrics: EER and WER, subjective metrics	✚ New metrics: GVD and pitch correlation 🟦 Metrics: EER and WER, subjective metrics	✚ New utility metric: UAR for SER 🟥 GVD, pitch correlation and subjective metrics

Figure 2.4: Evolution of VPC series from 2020 to 2024.

The initial VPC2020 [17] set the groundwork by focusing on anonymizing speaker identity without degrading speech intelligibility and naturalness. It introduced two baseline anonymization methods based on x-vectors [37] and neural waveform models [65], utilizing metrics such as WER and EER for evaluation. Participants were challenged to develop systems that could effectively transform speech to conceal the original speaker's identity while ensuring minimal performance loss in ASR tasks.

Building on this foundation, VPC2022 [63] introduced more sophisticated evaluation scenarios, including a semi-informed attack model where adversaries have access to anonymized enrollment data. This scenario presented a more realistic threat, compelling participants to design systems resilient to attackers with prior knowledge of the anonymization process. Additionally, VPC2022 expanded the evaluation framework with new utility metrics, such as pitch correlation (ρ_{F_0}) [65] and gain of voice distinctiveness (G_{VD}) [66, 67], to assess the preservation of paralinguistic features in anonymized speech. These enhancements underscored the delicate balance between privacy and utility, especially in multi-party conversational contexts.

The latest iteration, VPC2024 [24], further refined the challenge by simplifying the evaluation process and broadening the scope of utility metrics to include speech emotion recognition (SER) alongside ASR. This inclusion recognizes the importance of preserving emotional nuances in applications like healthcare and human-computer interaction [68]. VPC2024 also introduced advanced baseline systems utilizing neural audio codecs and sophisticated language modeling techniques, resulting in more natural and intelligible anonymized speech. The evaluation protocols and the deprioritization of voice distinctiveness and intonation preservation reflect a strategic shift towards practical applications where these attributes are less critical.

The VoicePrivacy Challenge series has made substantial contributions to the field of voice anonymization by providing a structured environment for benchmarking and innovation. Notable contributions include:

- **Development of Robust Anonymization Techniques:** Each VPC iteration has pushed the boundaries of voice anonymization, from x-vector-based methods in VPC2020 to the incorporation of neural audio codecs in VPC2024. These advancements have enhanced the robustness of anonymization systems against various attack models, including those involving semi-informed adversaries.

- **Establishment of Standardized Evaluation Frameworks:** VPC has established comprehensive evaluation protocols that integrate objective metrics (e.g., WER, EER) with subjective assessments (e.g., speech naturalness, intelligibility). This framework has been crucial in explaining the trade-offs between privacy and utility, identifying weaknesses in current systems, and guiding future improvements.

- **Expansion of Utility Metrics:** The introduction of SER as a utility metric in VPC2024 highlights the broader applicability of anonymized speech, particularly in domains requiring emotional expressiveness such as social robotics and mental health applications.

Despite these advancements, several research gaps remain. Achieving a robust anonymization that maintains high utility across diverse applications, especially under complex adversarial scenarios, continues to be challenging. Additionally, there is a need for more nuanced evaluation metrics that can effectively capture the intricacies of speaker identity concealment and the preservation of paralinguistic information in anonymized speech.

2.1.3 Attacker Models

In the context of voice anonymization, understanding the capabilities and objectives of potential attackers is crucial for developing robust privacy-preserving techniques [11, 69, 70]. The VoicePrivacy Challenge (VPC) series has systematically categorized attacker models to simulate realistic threat scenarios and evaluate the resilience of anonymization systems. This section provides a comprehensive overview of the different attacker models considered across VPC2020, VPC2022, and VPC2024.

Figure 2.5 illustrates various attacker scenarios, each representing a different level of knowledge and capability with respect to the speaker anonymization system. This categorization helps assess the robustness of anonymization techniques across multiple threat models:

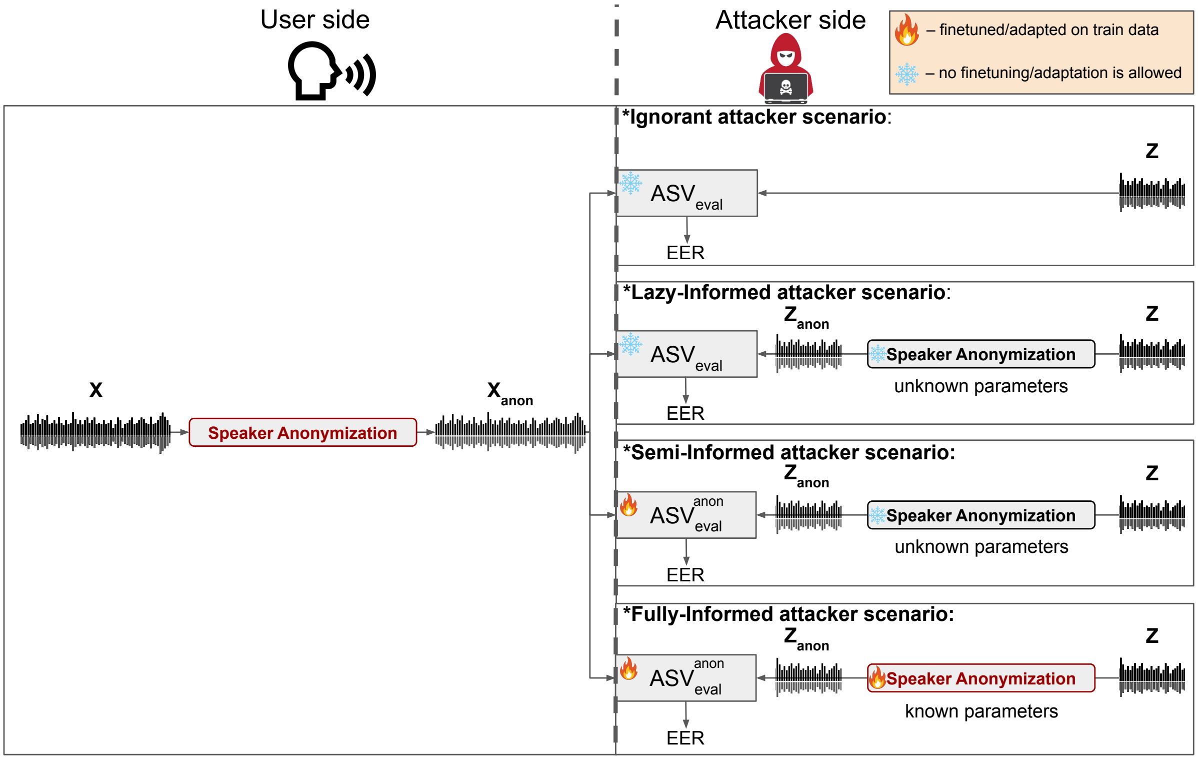


Figure 2.5: Different types of attacker scenarios.

- **Ignorant Attacker.** In this most basic threat model, the attacker (labeled as Ignorant Attacker) has no knowledge of the anonymization process and only has access to the anonymized test data ($\mathrm{X}_{\mathrm{anon}}$) and unanonimized enrollment data (Z) of target speaker. Without understanding the anonymization method or having access to any anonymized enrollment data, this attacker compares the anonymized test speech and unanonimized enrollment speech to attempt speaker recognition. The system's effectiveness here is measured through the (EER) achieved by the ASV attacker model ($\mathrm{ASV}_{\mathrm{eval}}$) in identifying speakers from anonymized speech.

- **Lazy-Informed Attacker.** The Lazy-Informed Attacker [71] has some awareness of the anonymization process but lacks specific parameter details. This attacker can anonymize enrollment data (Z) of the target speaker, which enables them to perform comparisons between anonymized test data (X_{anon}) and anonymized enrollment samples (Z_{anon}) of the target speaker. Although they do not know the precise parameters used for anonymization, they can try to adjust domain mismatches by anonymizing additional samples. This scenario evaluates the anonymization system's resistance against adversaries who have partial, but incomplete, information about the process. This attacker is not allowed to finetune ASV_{eval} model on anonymized train data.

- **Semi-Informed Attacker.** This Semi-Informed Attacker [70] is more sophisticated, with knowledge of the anonymization system and partial access to anonymization parameters, though not the exact values for each utterance. In addition, this attacker can fine-tune their ASV evaluation model ($\text{ASV}_{\text{eval}}^{\text{anon}}$) using anonymized training data to reduce domain mismatch between anonymized and non-anonymized speech, which strengthens their ability to identify speakers. This scenario tests the anonymization system against an adversary who uses both partial knowledge and training adaptation to increase attack effectiveness.
- **Informed Attacker.** The Fully-Informed Attacker [39] represents the most advanced threat, with complete knowledge of the anonymization system, including the exact parameters used for each anonymized utterance. This attacker can replicate the anonymization process with precise parameters, minimizing any domain mismatch between compromised and anonymized data. Additionally, the ASV evaluation model ($\text{ASV}_{\text{eval}}^{\text{anon}}$) can be fine-tuned to perform specifically against this known anonymization approach. This scenario allows a thorough evaluation of the anonymization system's vulnerabilities when facing an adversary with full access to the anonymization method and configurations.

Building on the attacker scenarios described, the effectiveness of an attack—whether successful or failed—is determined by the ability of the attacker to re-identify or link anonymized speech data back to the original speaker. A successful attack occurs when the anonymized speaker data is correctly identified as originating from the same speaker, demonstrating the anonymization system's vulnerability to re-identification. This success is measured by metrics like the EER, where a lower EER (as close to 0% as possible) means a more effective attack due to accurate speaker linking. Conversely, a failed attack means that the ASV system fails to associate the anonymized speech with the original speaker or results in incorrect matches, suggesting the anonymization system has effectively protected speaker privacy. High EER values (close to 50%), approaching levels expected from random guesses, indicate that the anonymized data is sufficiently unlinkable, rendering the attack ineffective. Thus, the EER metric serves as a crucial benchmark across the varying attacker models, providing insight into how well an anonymization system can withstand attempts at speaker re-identification by adversaries with different levels of knowledge and adaptation.

One more important fact to note is that there are two types of anonymization levels: speaker-level and utterance-level. In speaker-level anonymization, all utterances from a single speaker are consistently transformed using the same mapping. This ensures that while the original speaker's identity is concealed, the anonymized voice remains uniform across different utterances, allowing for recognition of the anonymized speaker in applications requiring speaker consistency. Conversely, utterance-level anonymization treats each utterance independently, applying different transformations to each one even if they originate from the same speaker. This results in higher privacy protection by preventing any linkage between utterances and the original speaker but may compromise applications that rely on speaker characteristics remaining consistent across utterances.

Privacy vs Utility Tradeoff

It is important to note that there is a tradeoff between privacy and utility metrics in evaluating speaker anonymization systems [72]. This tradeoff arises because, in speaker anonymization, two objectives must be balanced: maximizing privacy by effectively concealing speaker identity and maintaining the utility of the speech data for downstream tasks, such as ASR and SER.

As illustrated in Figure 2.6, privacy metrics focus on how well the anonymization conceals identifiable features. These include both objective measures, such as Equal Error Rate (EER) for verifiability, and subjective assessments that determine whether the anonymized speech can still be linked back to the original speaker. Additional privacy evaluations, like invertibility, assess the risk of reversing the anonymization to recover the original identity.

The evaluation process for VPC2024 employs four different EER conditions, each targeting specific privacy thresholds: 10–20%, 20–30%, 30–40%, and above 40%. These conditions are critical because they enable a comprehensive assessment of anonymization systems across different levels of privacy protection:

1. **EER**$_1$ (10–20%): Focuses on low-privacy requirements where utility is prioritized. Systems performing well here are ideal for scenarios where some speaker identity risk is tolerable.

2. **EER**$_2$ (20–30%): Represents balanced privacy-utility trade-offs suitable for moderately secure applications.

3. **EER$_3$** (30–40%): Evaluates systems with strong privacy requirements, where anonymization effectiveness outweighs minor losses in utility.

4. **EER$_4$** (Above 40%): Targets high-security use cases requiring maximum privacy, even at the expense of substantial utility reduction.

These conditions are essential for simulating real-world application scenarios. By evaluating systems within these thresholds, researchers can develop solutions tailored to specific needs, ranging from customer-facing services (low EER) to highly secure environments (high EER).

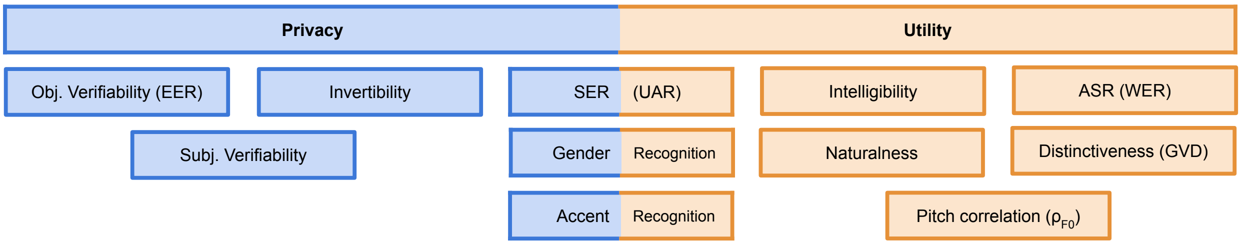


Figure 2.6: Tasks and metrics in the privacy-utility tradeoff for speaker anonymization, with privacy shown in blue (left half) and utility in orange (right half). Tasks that can be applied to both privacy and utility (e.g., gender recognition), depending on the requirements, are displayed in the middle.

On the other hand, utility metrics measure the preserved functionality of the anonymized speech in applications. For instance, ASR performance is evaluated by Word Error Rate (WER), while SER utility can be measured by Unweighted Average Recall (UAR). Other metrics like distinctiveness and pitch correlation also play roles in assessing utility, especially in terms of how well the speech retains natural and distinctive acoustic features post-anonymization.

Interestingly, some metrics, such as gender and accent recognition, serve dual purposes and can be applied to both privacy and utility assessments. Depending on the context, these metrics can either evaluate the anonymization system's success in hiding speaker characteristics (privacy) or confirm that anonymized speech still retains useful information (utility).

Achieving the ideal balance between privacy and utility is challenging; enhancing anonymization to secure privacy often leads to some loss in utility, as excessive concealment can degrade the speech's intelligibility or naturalness. This balancing act underscores the complexity of designing robust anonymization systems that meet the demands of privacy and usability in real-world scenarios.

2.1.4 Evaluation metrics

The evaluation framework for the VoicePrivacy Challenges is designed to assess the effectiveness of anonymization systems in terms of both privacy (the concealment of speaker identity) and utility (the preservation of speech intelligibility and usefulness for downstream tasks). Each edition of the challenge has introduced and refined various metrics to balance these concerns, reflecting the evolving requirements and complexities in voice anonymization research. Table 2.2 illustrates comparison of Evaluation Metrics across VPC2020, VPC2022, and VPC2024.

Table 2.2: Comparative Overview of Evaluation Metrics across VPC2020, VPC2022, and VPC2024

Metric	**VPC2020**	**VPC2022**	**VPC2024**
Anon. Level	Speaker	Speaker	Utterance
EER	Ignorant	Lazy/Semi-Informed	Semi-Informed
WER	✓	✓	✓
ρ_{F_0}	×	✓	×
GVD	×	✓	×
UAR	×	×	✓
Subj. Metrics	✓	✓	×

VPC2020 Evaluation Metrics

In its first edition, VPC2020 established foundational metrics to evaluate anonymization systems. Participants were required to submit both objective and subjective evaluation results. The primary privacy metric was the Equal Error Rate (EER), which serves as a critical measure in speaker verification systems. EER is defined as the point where the false acceptance rate (FAR) equals the false rejection rate (FRR). A higher EER indicates that the system is less accurate at verifying whether two speech samples are from the same speaker, implying better anonymization. The goal was to increase the EER, making it more difficult for an attacker to correctly identify speakers. In VPC2020, EER was evaluated under an ignorant attacker model, where the attacker has no prior knowledge of the anonymization process or access to anonymized data. The formula for EER can be found in Equation 2.1.

Complementing the privacy assessment, the Word Error Rate (WER) functioned as the main utility metric. WER measures the performance of automatic speech recognition (ASR) systems by calculating the percentage of words incorrectly recognized compared to the reference transcription. A lower WER indicates better preservation of linguistic content, ensuring that the anonymized speech remains useful for ASR applications. However, anonymization processes should not excessively increase WER, as it would degrade the utility of the speech data. This metric ensured that while speaker identity was concealed, the speech remained intelligible for downstream tasks. The formula for WER can be found in Equation 2.2.

Subjective metrics in the evaluation include: speaker verifiability, speech intelligibility, and speech naturalness. As illustrated in Figure 2.7, each of these metrics will be evaluated via a unified subjective test procedure carried out by the organizers.

In Figure 2.7, a Trial Utterance is passed either directly or via a Speaker Anonymization process, resulting in an Anon Trial Utterance. Evaluators will assess intelligibility and naturalness by listening to a single utterance at a time, while verifiability relies on comparing two utterances (the enrollment and the trial).

- **Naturalness.** For a given audio sample (original or anonymized), evaluators will assign a score from 1 ("totally unnatural") to 10 ("totally natural"). They will be asked to focus on the quality and presence of any audio degradation rather than speech content.

- **Intelligibility.** Evaluators will again listen to a single audio sample, original or anonymized, and rate from 1 ("totally unintelligible") to 10 ("totally intelligible"). They are instructed to focus on how comprehensible the speech content is.

- **Speaker Verifiability.** In this test, an original enrollment utterance is compared with a trial utterance, which may be original or anonymized and may come from the same or a different speaker. Evaluators will rate perceived voice similarity on a scale of 1 ("definitely different speakers") to 10 ("definitely the same speaker").

Evaluators are given the following scenario to guide their ratings:

> Please imagine that you are working at a TV or radio company. You wish to broadcast interviews of person X, but person X does not want to disclose their identity. Various automated anonymization tools are available, but some introduce severe artifacts or make speech less intelligible. Your task is to balance privacy (speaker anonymity) with the quality of the broadcast (naturalness and intelligibility).

Separate, detailed instructions are provided for each of the three subjective metrics. Evaluators are reminded of the scenario each time they begin one of the three specific tasks (naturalness, intelligibility, or speaker verifiability).

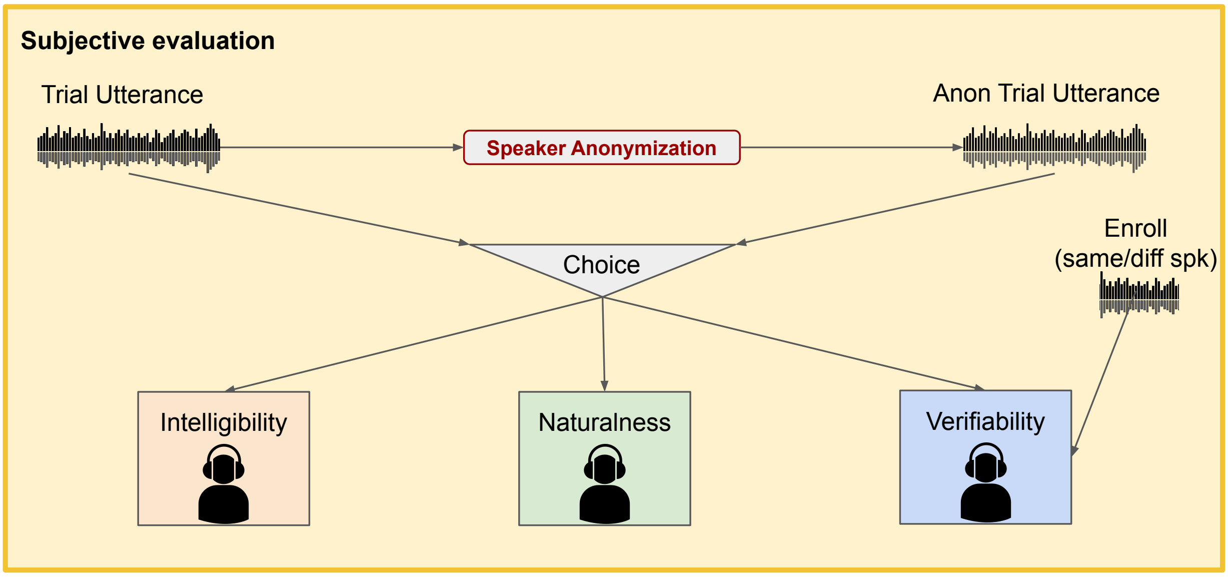


Figure 2.7: Evaluation pipeline for Subjective Metrics

VPC2022 Evaluation Metrics

Building upon the initial framework, VPC2022 introduced more nuanced metrics to address complex real-world scenarios. The Equal Error Rate (EER) was enhanced by calculating it under a semi-informed attacker model, where the attacker has access to anonymized enrollment utterances of the target speakers. This advancement simulated a more realistic threat scenario, emphasizing the need for anonymization systems to be robust against adversaries with partial knowledge of the anonymization process. The objective was to maintain a high EER even when attackers had some level of access to anonymized data.

The Word Error Rate (WER) remained the primary utility metric, ensuring that anonymization techniques did not excessively degrade the linguistic content of the speech, thus maintaining utility for ASR and other applications.

To assess the preservation of paralinguistic features, VPC2022 introduced the Pitch Correlation (ρ_{F_0}) (Equation 2.3) as a secondary utility metric. This metric is calculated as the correlation coefficient between the fundamental frequency contours of the original (P) and anonymized (Q) speech. Pitch is crucial for conveying prosody, intonation, and emotional cues. Preserving pitch enhances the naturalness and expressiveness of speech, which is important for applications beyond ASR. By measuring how well the anonymized speech preserved pitch information, the challenge encouraged participants to develop systems that maintained the quality and usability of speech data.

$$\rho_{F_0} = \frac{\sum_{t=1}^{T}(P_t - \bar{P})(Q_t - \bar{Q})}{\sqrt{\sum_{i=t}^{T}(P_t - \bar{P})^2}\sqrt{\sum_{t=1}^{T}(Q_t - \bar{Q})^2}}, \tag{2.3}$$

where $\bar{P}$ and $\bar{Q}$ are the averages of pitch countours of original and anonymized speech segments respectively.

Another significant addition was the Gain of Voice Distinctiveness (G_{VD}), which measured the distinctiveness of anonymized voices in multi-party conversations. This metric assessed whether anonymized voices of different speakers remained distinguishable from one another. While the primary goal is to anonymize speaker identity, it is important that different speakers do not become indistinguishable, as this could hinder communication in conversational scenarios. G_{VD} ensured that anonymization systems balanced privacy with the practical need for speaker differentiation in multi-speaker contexts.

In order to calculate G_{VD} (Equation 2.6), we need to introduce similarity matrix $S(i,j)$ (Equation 2.4) and diagonal dominance $D_{\text{diag}}(S)$ (Equation 2.5). First, similarity matrix $S(i,j)$ is calculated using log-likelihood ratio $\text{LLR}(x_k^{(i)}, x_l^{(j)})$, where $x_k^{(i)}$ – k-th segment from speaker i; n_j and n_i refer to the number of segments available for each speaker. Using this approach, two similarity matrices are computed: S_{orig} for the original speech and S_{anon} for the anonymized speech.

$$S(i,j) = \text{sigmoid}\left(\frac{1}{n_i n_j} \sum_{\substack{1 \le k \le n_i \\ 1 \le l \le n_j \\ k \ne l \text{ if } i=j}} \text{LLR}(x_k^{(i)}, x_l^{(j)}) \right) \tag{2.4}$$

Next, diagonal dominance is defined as the absolute difference between the mean values of the diagonal elements (representing intra-speaker similarity) and off-diagonal elements (representing inter-speaker similarity).

$$D_{\text{diag}}(S) = \left| \sum_{1 \le i \le N} \frac{S(i,i)}{N} - \sum_{\substack{1 \le j \le N \\ j \ne k}} \sum_{1 \le k \le N} \frac{S(j,k)}{N(N-1)} \right| \tag{2.5}$$

The gain of voice distinctiveness, G_{VD} is defined as the logarithmic ratio of the diagonal dominance for anonymized and original data matrices.

$$G_{\text{VD}} = 10 \log_{10} \frac{D_{\text{diag}}(S_{\text{anon}})}{D_{\text{diag}}(S_{\text{orig}})} \tag{2.6}$$

VPC2024 Evaluation Metrics

VPC2024 further refined the evaluation framework by introducing new metrics and simplifying the evaluation process to reflect emerging applications and practical considerations. The Equal Error Rate (EER) was tested under scenarios involving more advanced attacker models, including attackers using deep learning systems trained on anonymized data. This enhancement challenged participants to design anonymization systems that could withstand sophisticated attacks, emphasizing robustness and security.

The Word Error Rate (WER) continued to be a critical utility metric, with an emphasis on ensuring that anonymized speech retained sufficient linguistic content for ASR and other speech-based applications, such as natural language understanding and translation.

Recognizing the importance of emotional expressiveness in speech, VPC2024 introduced Speech Emotion Recognition (SER) Performance as a utility metric. This addition evaluated how well anonymized speech preserved emotional cues, which are critical for applications in healthcare, social robotics, and customer service. Preserving emotion enhances user interaction and the effectiveness of speech-based systems. The goal was to maintain high SER performance even on anonymized speech, demonstrating that the emotional content remained intact. In order to do that, Unweighted Average Recall (UAR) is used. The formula can be found in Equation 2.7.

$$UAR = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i}, \tag{2.7}$$

where C is the total number of emotion classes, TP_i, FN_i are amount of True Positives and False Negative accordingly for i-th class. removed to streamline the participation process.

The VoicePrivacy Challenge series has progressively refined its evaluation metrics to address the multifaceted nature of voice anonymization. By balancing privacy and utility concerns and introducing metrics that reflect real-world application needs, the challenges have guided the development of more robust and effective anonymization systems. Table 2.2 shows summary of Evaluation Metrics across VPC2020, VPC2022, and VPC2024.

2.1.5 Datasets and Corpora

The VoicePrivacy 2024 Challenge utilizes a variety of publicly available datasets for training, development, and evaluation purposes. These datasets provide the necessary resources to develop and benchmark voice anonymization systems, allowing for a comprehensive evaluation of both privacy protection and utility preservation. The datasets are carefully selected to cover a wide range of speech characteristics, speaker demographics, and paralinguistic information, ensuring that anonymization systems are tested in diverse real-world conditions.

Training Data

The training data provided for participants include speech datasets from different domains, such as audiobooks, conversations, and emotional speech recordings. These datasets are used to train both the anonymization models and the feature extractors (e.g., ASR, emotion recognition systems) needed to evaluate the systems. The participants can also use pre-trained models from various sources such as WavLM, Whisper, and wav2vec2, which are derived from these datasets to further improve the quality of their systems. Full set of allowed training data for VPC2024 is listed in Table 2.3.

Table 2.3: Summary of Training Datasets and Corpora Used in VoicePrivacy 2024 Challenge

Dataset	Main Purpose	Domain	Size/Hours	Description
LibriSpeech (train) [44]	ASR	Audiobooks	960 hours	Large-scale corpus of read English speech from over 2,300 speakers, used for ASR model training and anonymization evaluation.
Libri-light [73]	ASR	Audiobooks	60000 hours	A large-scale subset of LibriSpeech with unlabeled data, often used for unsupervised ASR.
CMU-MOSEI [74]	ASR	Multi-domain	23,500 videos	Multimodal dataset for emotion recognition.
VoxCeleb1 & 2 [33]	ASV	Online videos	1.2 mil utts	Speech extracted from video content, representing diverse accents and demographics for training speaker recognition.
RAVDESS [75]	SER	Emotions	24 speakers	Emotional speech and song database with calm, happy, sad, angry, fearful, surprise, and disgust expressions. Available in audio, video, and audiovisual formats.
MSP-Podcast [76]	SER	Podcasts	237 hours	A collection of podcast speech clips covering a range of emotions and natural conversational styles, used primarily for emotion recognition tasks.
VGAF [77]	SER	Emotions	120 hours	Video Gesture Analysis Framework dataset with vocal emotions
ESD [78]	SER	Emotions	175 hours	Emotional Speech Database with 350 utterances from 20 speakers in 5 emotions, enabling voice conversion research.
CREMA-D [79]	SER	Emotions	7442 utts	91 actors, 6 emotions, crowd-rated for emotion and intensity.
SAVEE [80]	SER	Emotions	480 utts	4 native English speakers with 7 emotion categories.
EMO-DB [81]	SER	Emotions	535 utts	German emotional database with 7 emotions, 10 speakers.
LibriTTS [82]	TTS	Audiobooks	585 hours	A dataset of English speech designed for text-to-speech synthesis tasks.
LJSpeech [83]	TTS	Audiobooks	24 hours	High-quality single-speaker dataset for TTS development, useful for voice conversion tasks.
VCTK [84]	VC	Read Speech	44 hours	Corpus of English speech from multiple accents, commonly used for ASR, TTS and VC.
MUSAN [85]	AUG	Misc	109 hours	Collection of music, speech, and noise samples for data augmentation.
RIR [86]	AUG	Room Impulse	900 RIRs	Room impulse response dataset for simulating reverberation.

Development and Evaluation Data

The development and evaluation sets include both anonymized and original speech samples. These sets are fixed and are used to assess the performance of the anonymization systems using the provided evaluation scripts. The main datasets used for evaluation are:

LibriSpeech (dev/test) [44]: This corpus contains English speech sampled at 16 kHz, sourced from audiobooks, making it well-suited for ASR research. It will serve as the primary dataset for ASV and ASR evaluations. The development and evaluation subsets of LibriSpeech remain consistent with previous challenge editions. Statistics for this dataset are described in Table 2.4.

Subset			**Female**	**Male**	**Total**	**#Utterances**
Development	LibriSpeech dev-clean	Enrollment	15	14	29	343
		Trial	20	20	40	1,978
Evaluation	LibriSpeech test-clean	Enrollment	16	13	29	438
		Trial	20	20	40	1,496

Table 2.4: Number of Speakers and Utterances for ASR and ASV development and evaluation.

IEMOCAP [27]: This dataset includes 12 hours of audio-visual emotional speech sampled at 16 kHz, with recordings of two-speaker dialogues involving five female and five male English-speaking actors. Four emotions are focused on for evaluation: neutral, sadness, anger, and happiness. To balance the class distribution, the happiness and excitement categories are merged. Due to the limited number of speakers and total data size, a leave-one-conversation-out cross-validation approach is applied. Within each cross-validation fold, conversations from eight speakers are allocated for training the SER evaluation model, while the remaining conversations from two speakers are designated for development and evaluation purposes. Statistics for this dataset are described in Table 2.5.

	Session 1	**Session 2**	**Session 3**	**Session 4**	**Session 5**
Female	528	481	522	528	590
Male	557	542	629	503	651

Table 2.5: Number of utterances per gender in each session of IEMOCAP.

2.1.6 VPC2024 Baseline Models

In this subsection we introduce baseline methods from VPC2024 as it has the strongest systems in comparison to VPC2020 and VPC2020. The VoicePrivacy 2024 Challenge provides 6 baseline anonymization systems (B1-B6). These baselines represent different approaches to anonymizing speaker identity while preserving the linguistic content and utility of speech. Below is a brief description of each system.

B1: Anonymization using x-vectors and a Neural Source-Filter Model.

Baseline B1 [37] is based on the x-vector-based anonymization method, which was also used in previous challenges (VPC2020 and VPC2022). The system is trained on LibriSpeech and VoxCeleb datasets for feature extraction and synthesis. It is illustrated on Figure 2.8 and involves three main steps:

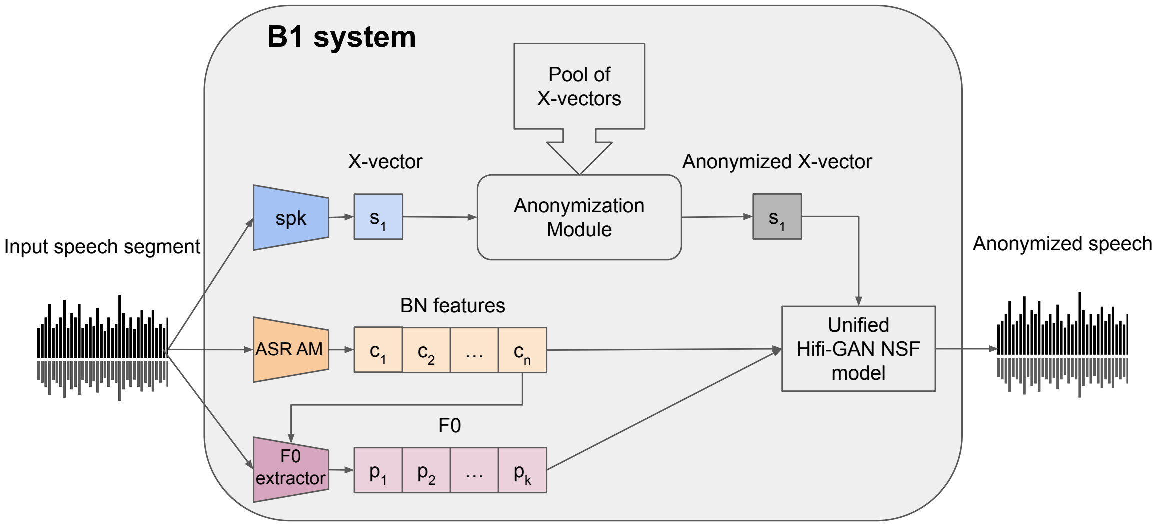


Figure 2.8: B1 pipeline

- **Feature Extraction:** Speaker x-vectors (spk), bottleneck (BN) features (ASR AM), and fundamental frequency (F0 extractor) are extracted from the input speech [87, 88, 89].
- **X-vector Anonymization (Anonymization Module):** The x-vectors, which encode speaker-specific information, are replaced with pseudonymous speaker x-vectors chosen from a pool of external speakers [36].
- **Speech Synthesis:** The anonymized x-vector, along with the original BN and F0 features, is passed through a neural source-filter (NSF) model to generate anonymized speech [65, 90].

B2: Anonymization using McAdams Coefficient.

Baseline B2 [35] uses a signal processing-based method, relying on linear predictive coding (LPC) and the McAdams coefficient [91] to anonymize speech. This method is computationally lightweight and does not require training data. The pipeline is shown on Figure 2.9.

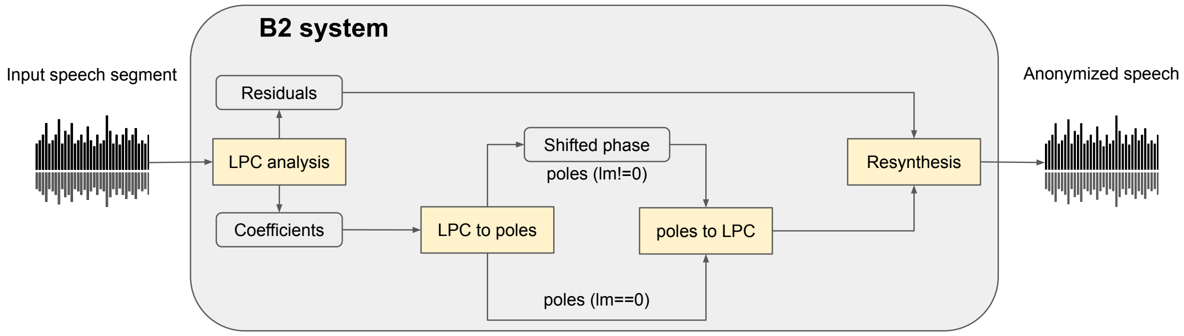


Figure 2.9: B2 pipeline

- **LPC Analysis:** The input speech is analyzed using LPC to extract the pole positions representing spectral peaks (formants).

- **McAdams Transformation:** The phases of the poles are transformed using the McAdams coefficient to shift formant frequencies, thereby changing the perceived speaker identity while retaining the overall structure of the speech.

- **Resynthesis:** The transformed LPC coefficients are combined with the original residual signal to synthesize anonymized speech.

B3: Anonymization using Phonetic Transcriptions and GAN

Baseline B3 [92] incorporates a generative adversarial network (GAN) to generate pseudonymous speaker embeddings. The system uses datasets such as LibriTTS, RAVDESS, and ESD for training. The pipeline is illustated on Figure 2.10.

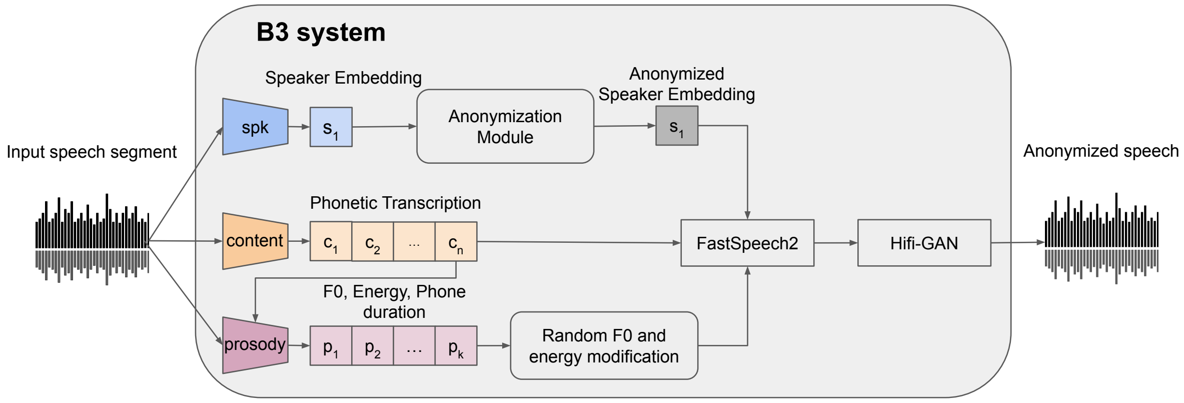


Figure 2.10: B3 pipeline

- **Feature Extraction:** Extracts speaker embeddings (spk) [93], phonetic transcription (content) [94, 95], F0, energy, and phone durations (prosody) from the original speech.

- **GAN-based Speaker Anonymization:** A Wasserstein GAN (Anonymization Module) [96] generates a new speaker embedding, replacing the original embedding. F0 and energy values are randomly modified to remove individual prosody patterns.

- **Speech Synthesis:** The anonymized features and phonetic transcription are fed into a FastSpeech2-based synthesis model [97] to generate anonymized spectrogram and then to HiFi-GAN for speech generation [98].

B4: Anonymization using Neural Audio Codec (NAC) Language Modeling

Baseline B4 [38] uses neural audio codecs and language models to anonymize speech. The method is illustrated on Figure 2.11.

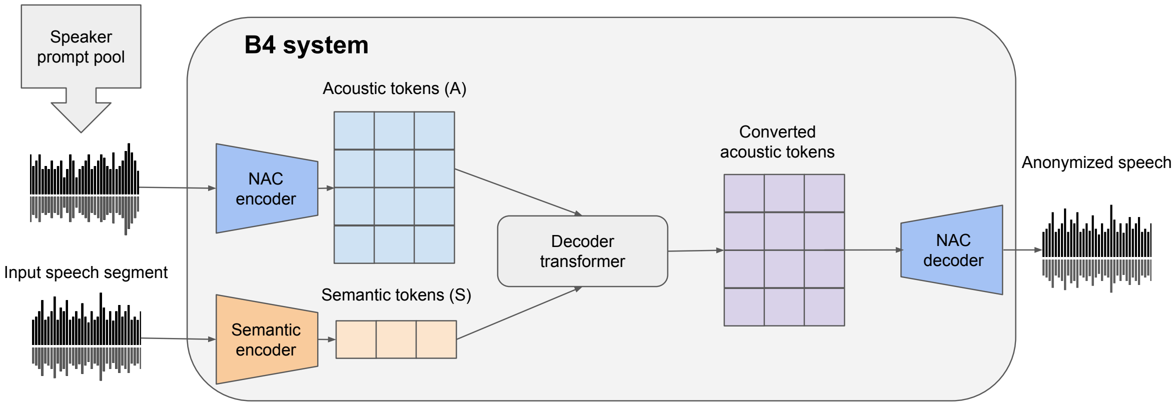


Figure 2.11: B4 pipeline

- **Semantic Extraction:** HuBERT [99] extracts semantic tokens from the input speech to represent linguistic content. The output is represented as a sequence of integers $S \in s_1, ..., s_{T_s}$, where T_s denotes the number of frames, and each integer corresponds to a codeword index.

- **Acoustic Token Generation:** Acoustic tokens from a pool of pseudonymous speakers are concatenated with the semantic tokens. The output is represented as N sequences of integers $A \in \{A_1, ..., A_N\}^{N\times T_a}$, where T_a is the number of frames and where each integer is a codeword index to corresponing codebooks

- **Speech Generation:** A GPT-like transformer generates new acoustic tokens, which are then decoded by a neural audio codec to synthesize anonymized speech.

This approach leverages models like EnCodec [100] and Bark to handle both semantic and acoustic aspects of the speech.

B5: Anonymization using ASR-Bottleneck with Vector Quantization (VQ)

Baseline B5 [39] improves on the x-vector-based methods by introducing vector quantization (VQ) to enhance the separation of linguistic and speaker information. The training data includes LibriSpeech and VoxPopuli datasets.

- **Feature Extraction:** Uses an ASR acoustic model (ASR AM) to extract VQ-bottleneck features that represent the linguistic content of the speech.
- **VQ-based Anonymization:** The VQ process replaces continuous feature vectors with a finite set of vectors to reduce the amount of speaker-specific information.
- **Speech Synthesis:** A HiFi-GAN synthesizer generates anonymized speech from the VQ-bottleneck features, fundamental frequency (F0), and a target one-hot speaker vector.

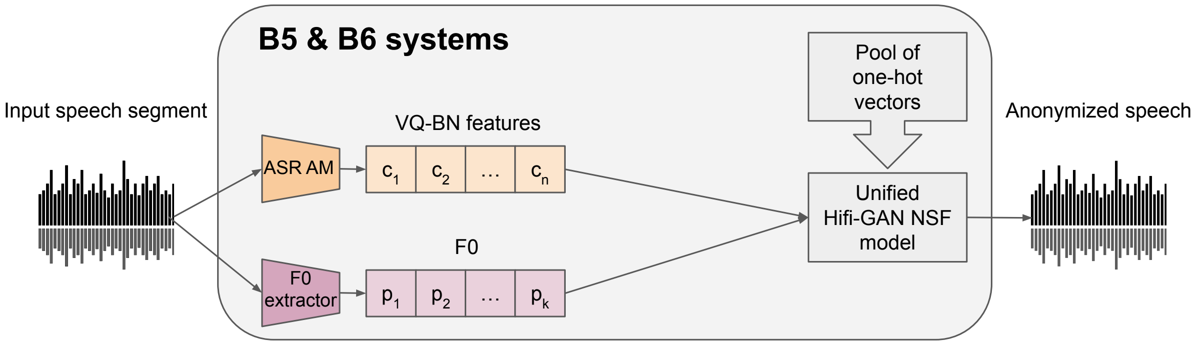


Figure 2.12: B5 and B6 pipelines

B6: Anonymization using ASR-BN and TDNN-F

Baseline B6 [39] is similar to B5 but uses a different ASR acoustic model for bottleneck feature extraction. The pipelines B5 and B6 are shown on Figure 2.12.

- **ASR-BN Extraction (ASM AM):** The ASR model consists solely of time-delay neural network factorized (TDNN-F) layers, unlike B5 which uses wav2vec2 pre-training.
- **VQ-based Anonymization:** As in B5, VQ is applied to the bottleneck features to ensure better separation between linguistic and speaker attributes.

- **Speech Synthesis:** The anonymized features are used to synthesize speech using the HiFi-GAN model.

Both B5 and B6 focus on improving privacy protection by leveraging vector quantization to reduce the amount of speaker-identifying information in the bottleneck features.

2.2 Connection of Speaker Anonymization Approaches to Voice Conversion and TTS

Speaker anonymization, voice conversion (VC) [38, 101], and text-to-speech (TTS) [40, 41, 60] synthesis are interlinked fields that leverage advanced signal processing and machine learning techniques to manipulate speech data while balancing privacy and intelligibility. While speaker anonymization aims to mask speaker identity, VC and TTS focus on altering speech characteristics or generating speech from textual inputs, respectively. This section explores how these domains intersect, particularly how methods and insights from VC and TTS have been adapted to advance speaker anonymization. Furthermore, it incorporates a detailed discussion of their differences and similarities, as illustrated in Figure 2.12.

Figure 2.13 visually highlights the differences and similarities between Speaker Anonymization and Voice Conversion systems. The key distinctions and overlaps are as follows:

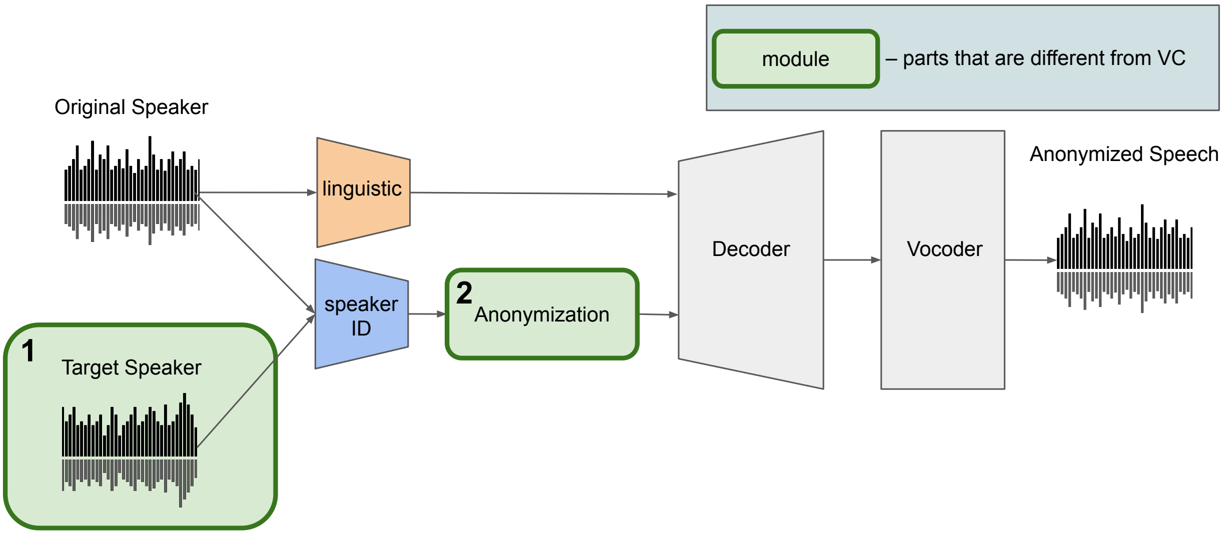


Figure 2.13: Comparison between Speaker Anonymization and Voice Conversion. The differences are highlighted in green.

- **Target Speaker (Difference 1):** In Voice Conversion, setting a target speaker is a fundamental requirement. The system transforms the original speaker's characteristics to match those of a predefined target speaker.

Conversely, Speaker Anonymization does not necessitate a target speaker. Anonymization can be performed by manipulating the features of the original speaker to hide identity without mapping to another speaker.

- **Anonymization Module (Difference 2):**

  Speaker Anonymization systems include a dedicated anonymization module to ensure the speaker's identity is effectively concealed. This module is not present in Voice Conversion systems, which focus solely on altering speaker characteristics without anonymization.

- **Shared Components (Similarities):**

  Both systems rely on similar foundational components, such as linguistic feature extraction, decoding, and vocoders, to reconstruct the speech. These shared modules highlight the overlap in methodologies and underline the mutual reliance on advanced speech synthesis techniques.

The integration of VC and TTS techniques into speaker anonymization is not without challenges. Ensuring robustness against adversarial attacks, such as informed attackers with access to TTS pipelines, remains a critical area of research. Additionally, while TTS and VC prioritize intelligibility and naturalness, anonymization systems must often sacrifice these attributes to ensure higher privacy levels.

2.3 Summary of the chapter

In summary, the literature review on speaker anonymization and disentanglement learning reveals a dynamic field aimed at balancing privacy and utility in speech data. Through initiatives like the VPC, standardized benchmarks have emerged, driving improvements in anonymization techniques.

Chapter 3

Disentanglement-based Approaches for Anonymization

In this chapter, we cover disentanglement learning for speech processing, which separates linguistic and paralinguistic features to help anonymization, and provide details on various disentanglement methods such as VAEs [102] and Neural Codecs [29], along with their applications in speaker verification, emotion recognition, and voice conversion. Next, we describe our submissions for the Voice Privacy Challenge 2024. Rather than proposing one novel speech anonymization system, we enhance the provided baselines to meet all required conditions and improve evaluated metrics. Specifically, we implement emotion embedding and experiment with WavLM and ECAPA2 speaker embedders for the B3 baseline. Additionally, we compare different speaker and prosody anonymization techniques. Furthermore, we introduce Mean Reversion F0 for B5, which helps to enhance privacy without a loss in utility. Finally, we explore disentanglement models, namely ß-VAE and NaturalSpeech3 FACodec.

Given the baselines provided by organizers, we test various techniques and methods aimed at improving the evaluated metrics. Specifically, we create submissions for all four of EER conditions by using modified NaturalSpeech3 FACodec for condition EER_1, a modified version of B3 [92] for conditions EER_1 and EER_2, and enhanced B5 [39] for EER_3 and EER_4.

The rest of the chapter is constructed as follows: Section 3.1 provides overview of disentanglement-based approaches Section 3.2 summarizes the two baselines that we used for our submissions, introduces all the modifications and proposed techniques used in our experiments, provides the detailed results and the systems that we submitted for evaluation, and concludes our findings.

3.1 Disentanglement Learning in Speech

3.1.1 Overview of Disentanglement Learning in Speech

Disentanglement learning has gained attention in machine learning [103, 104, 105, 106, 107], especially within the realm of speech processing [108], due to its capacity to break down complex data into independent factors of variation. Specifically, disentanglement aims to separate data into meaningful components, each representing distinct characteristics of the input. In the context of speech processing, disentangled representations can isolate various features such as speaker identity, linguistic content, prosody, and emotion. This capability is particularly valuable for applications that require both privacy and usability, such as speech privacy protection and anonymization, where it is essential to handle sensitive information selectively.

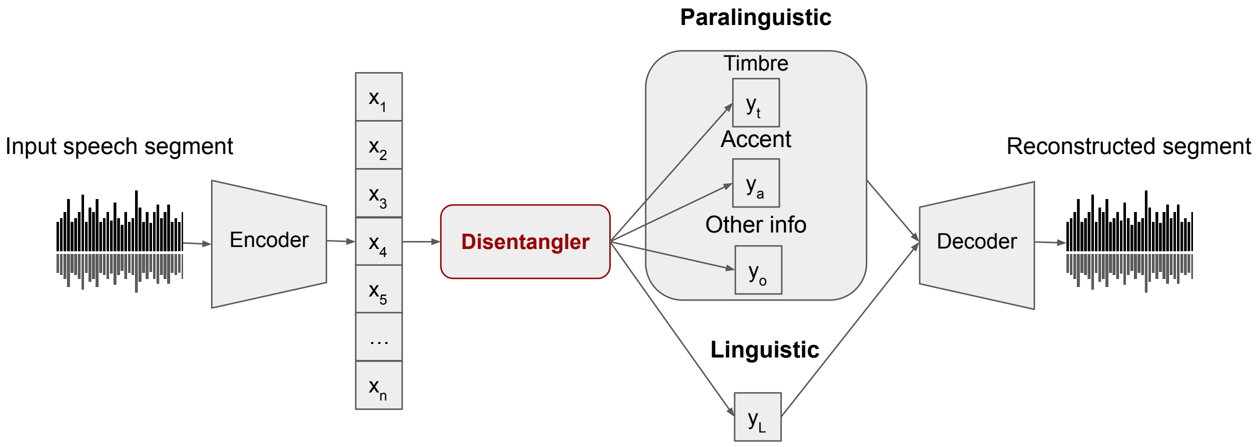


Figure 3.1: Training Disentanglement Pipeline

The diagram in Figure 3.1 illustrates a typical training pipeline for disentanglement learning in speech processing. The process begins with an input speech segment fed into an Encoder, which captures and transforms the audio data into feature representations (denoted by $x_1, x_2, x_3, ..., x_n$). These representations are then processed by a *Disentangler*, which separates the features into two main categories: paralinguistic and linguistic components.

- **Paralinguistic features** (top of the diagram) represent non-linguistic aspects of speech, including timbre, accent, and other attributes that may carry speaker-specific information. These components are denoted by variables y_t, y_a, y_o that correspond to timbre, accent and other information respectively [106, 107]. It encapsulates qualities that contribute to the speaker's identity and expressive style but are not tied to the actual linguistic content of the speech [108]. It is important to note that timbre and accent are shown as examples, and different disentanglement systems may separate other paralinguistic features, such as age and emotions, etc.

- **Linguistic features** (bottom of the diagram) contain the spoken content, labeled as y_L. This category represents the information necessary to understand the words and meaning, separated from the speaker's identity and expressive characteristics.

After disentanglement, the separated components are passed to a Decoder, which attempts to reconstruct the original speech segment using the distinct paralinguistic and linguistic factors. This reconstructed segment maintains the intelligibility and content of the original speech while allowing for modifications to protect speaker identity or other sensitive information.

Disentangled representations in speech processing can significantly enhance privacy by controlling which speech attributes are preserved or masked. For example, by concealing speaker timbre while retaining the spoken content, users can engage in private communication or interact with voice-activated systems without compromising personal information. Moreover, disentanglement learning aligns with broader machine learning goals of interpretability and modularity, enabling a structured approach to data representation.

3.1.2 Methods and Approaches for Disentanglement in Speech

Several approaches have been developed to disentangle latent representations of speech. These methods fall broadly into different categories: from unsupervised to supervised approaches. In unsupervised learning, models such as Variational Autoencoders (VAEs) [102, 109] and Generative Adversarial Networks (GANs) [110] are widely employed. These models attempt to uncover hidden factors in the data without labeled training data, making them suitable for privacy-focused tasks where obtaining labeled datasets may be impractical.

Williams' work [111] demonstrates how end-to-end self- and semi-supervised methods can effectively disentangle speech signals for specific tasks such as speaker diarization or privacy masking, where certain attributes (e.g., speaker identity) are isolated while preserving others (e.g., content). These architectures rely on deep learning techniques like latent variable models, where different aspects of speech, such as speaker identity and spoken content, are encoded into separate latent spaces.

In supervised learning, disentanglement is guided by explicitly defined labels or constraints on the data. While this approach provides more control over the factors of interest (e.g., identity vs. content), it often requires larger datasets and more detailed annotation. Neural Codecs correspond to these types of approaches [112, 29]

In this work, we focus on VAE-based [102] and Neural Codec-based [29] approaches. Firstly, VAE-based approaches are designed to address a significant challenge in disentanglement learning: how to make sure that the learned latent representations remain consistent when exposed to changes in visual or audio input factors. The key components of this method are KL divergency and Mutual Information losses, which ensure that the model maintains consistent latent representations despite variations in input data. For instance, when the input data involves speech recordings with slight differences in pitch, speed, or background noise, the latent factors representing speaker identity should remain stable. This is crucial for speech anonymization applications, where you want to modify or anonymize specific factors like speaker identity while ensuring the consistency of other variables such as spoken content.

Secondly, Neural Codec-based approaches use gradient reversal layers (GRL) [113], teacher distillation [112], and vector quantization [114] to enhance speech quality and disentanglement. GRL helps the model learn representations invariant to factors like speaker identity by reversing gradients, enabling better separation of speaker characteristics from linguistic content. This is crucial for tasks like voice anonymization, where consistent separation of content and identity is required. Teacher distillation improves fidelity by training the model to replicate high-quality features from a pre-trained "teacher" model. Combined with vector quantization, which discretizes the latent space, these methods ensure the codec captures critical linguistic characteristics efficiently. This enables fine control over audio features, maintaining quality and naturalness in modified outputs.

3.1.3 Applications of Disentangled Speech Representations

This subsection illustrates how disentanglement learning can be applied to various downstream tasks. Figure 3.2 presents the general concept, showing how utterances from two speakers (Y and Z) are processed through an Encoder and Disentangler to obtain paralinguistic representations $(y_t, y_a, y_o, z_t, z_a, z_o)$ and linguistic representations (y_L, z_L) for each speaker. Specifically, **Paralinguistic representations** capture attributes unrelated to the linguistic content, such as timbre (y_t, z_t) and accent (y_a, z_a), along with other paralinguistic information (y_o, z_o). **Linguistic representations** (y_L, z_L) represent the actual spoken content of each speaker, independent of their identity or expressive style.

Once these representations are extracted, they can be used for a variety of tasks:

- For **ASV (Automatic Speaker Verification)**, the timbre-related components y_t, z_t offer essential speaker identity cues.

- **SER (Speech Emotion Recognition)** can utilize other paralinguistic elements, such as y_o, z_o, which contain expressive characteristics relevant to emotion detection.

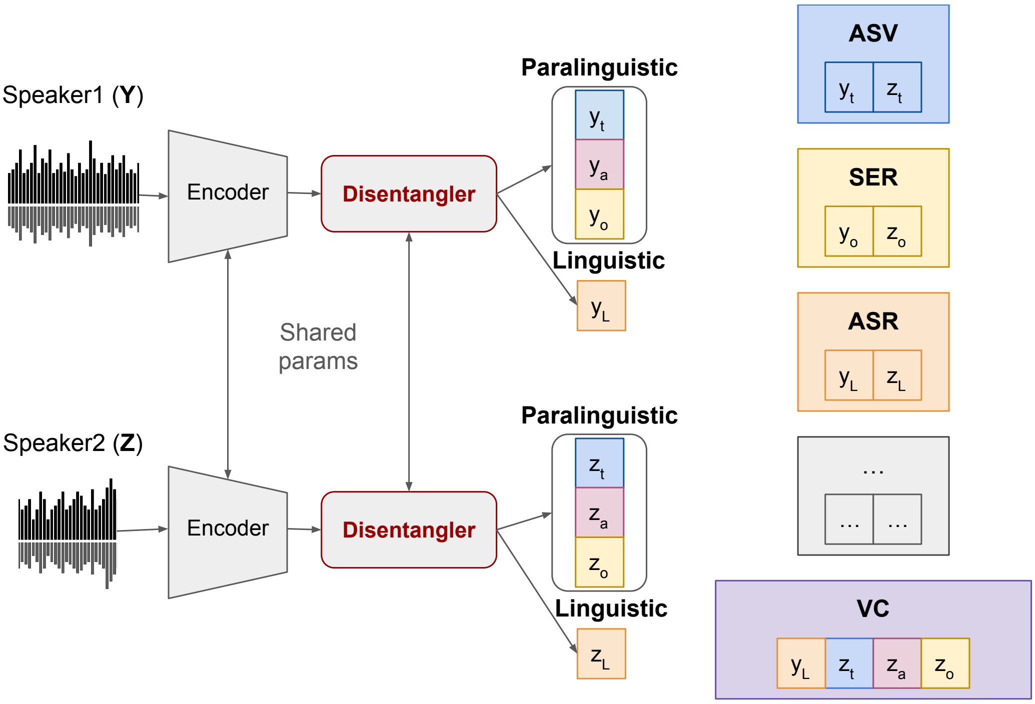


Figure 3.2: Applications of Speaker Disentanglement Representations

- For **ASR (Automatic Speech Recognition)**, linguistic representations y_L, z_L are critical, as they capture the content of the speech necessary for accurate transcription.

- For **Voice Conversion (VC)**, by combining the linguistic content from Speaker1 (y_L) with the paralinguistic features (such as timbre and accent) from Speaker2 (z_t, z_a, z_o), we can synthesize speech that retains the content of Speaker1 but with the voice characteristics of Speaker2.

It is important to note that the paralinguistic aspects shown here (e.g., timbre and accent) are provided as examples. Different disentanglement models may capture other attributes, such as age, emotions, speaker identity, or various acoustic details, depending on the target application. The flexibility of disentangled representations enables a modular approach to feature control, making it possible to tailor anonymization and conversion models to meet specific requirements across multiple tasks.

3.1.4 Two recent approaches for Disentanglement Learning.

In this section we overview two models based on disentanglement which can be adapted for Speaker Anonymization use: β-VAE [102] and NaturalSpeech3 FACodec [29].

β-VAE.

To begin with, we describe β-VAE model. The main concept of the Beta Variational Autoencoder (β-VAE) model is to disentangle speech representations to separate content from speaker identity, particularly for cross-lingual voice conversion tasks. By leveraging the structure of β-VAE, the model includes two encoders that independently learn to represent content and speaker information. The content encoder captures frame-level features, which are time-variant aspects of the speech, ensuring that the spoken content is accurately represented across frames. Meanwhile, the speaker encoder generates a single, time-invariant vector for each utterance, effectively isolating the speaker's identity. This separation allows for flexible conversion where content from one language can be combined with speaker identity from another, enabling one-shot cross-lingual voice conversion. The training pipeline of β-VAE is illustrated on Figure 3.3.

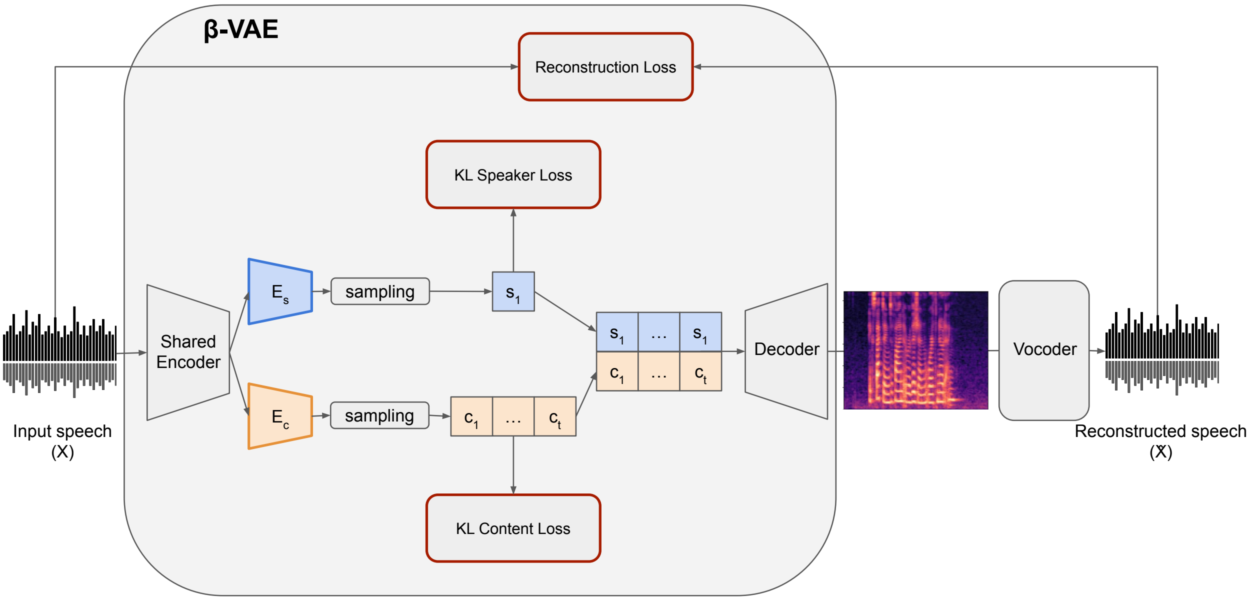


Figure 3.3: β-VAE training pipeline.

To ensure accurate disentanglement of content and speaker identity, the model uses two weighting parameters, β_s and β_c, applied to the Kullback-Leibler (KL) divergence terms in the loss function. These parameters control the information retained in each encoder, acting as "gates" to regulate the data flow through each representation. The model's overall loss function consists of three parts: reconstruction loss, content KL divergence, and speaker KL divergence. The loss function can be represented as:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \beta_c \mathcal{L}_c + \beta_s \mathcal{L}_s, \tag{3.1}$$

where:

- **The reconstruction loss term $\mathcal{L}_{\text{rec}}$** minimizes the difference between the original and reconstructed speech, ensuring that both the content and the identity of the speaker are preserved in the output.

- **The content KL divergence term** $\mathcal{L}_c$ regulates the content encoder by controlling the information flow in the content representation $c_1, ..., c_t \in \mathbb{R}^d$, where d is a vectors' dimension.

- **The speaker KL divergence term** $\mathcal{L}_s$ limits the speaker encoder's ability to capture speaker identity information through $s_1 \in \mathbb{R}^D$, where D is a vector's dimension.

The model's architecture includes a decoder that combines the outputs from both encoders to reconstruct speech with the source's content and the target's speaker identity. Here, $c_1, ..., c_t$ and s_1 are latent variables that separately encode content and speaker information. Because the speaker's identity (s_1) is represented at the utterance level, it is duplicated t times to align with the content before concatenation. The reconstruction loss $\mathcal{L}_{\text{rec}}$ encourages both $c_1, ..., c_t$ and s_1 to retain maximal information about the speech signal, while the KL divergence terms with weights β_c and β_s ensure that only relevant information is captured by each representation.

To achieve the complementary nature of these representations, the KL divergence terms act as regularizers. By choosing appropriate values for β_c and β_s, the model ensures a compact representation that prevents overlap between content and speaker information. The parameters β_c and β_s are set by tuning for optimal performance, and their absolute values help maintain complementary, speaker-independent content and identity information in a voice conversion.

Speaker and content encoder consist of 1D convolution blocks and ReLU activations with dropouts. Speaker encoder has additional fully-connected (FC) layer and global-average pooling to get utterance-level representation. In contrast, content encoder does not have pooling, but has two self-attention blocks to capture frame-level representations instead. Decoder consists of convolutional layers with self-attention blocks similarly to content encoder. Additionally, it has 5 Layers of 1D convolutional blocks and dropout.

NaturalSpeech3 FACodec.

Next, we introduce NaturalSpeech3 FACodec. The FACodec module in NaturalSpeech 3 utilizes a carefully structured pipeline, as illustrated in the diagram on Figure 3.4, to decompose speech into distinct, disentangled attributes for zero-shot speech synthesis. By factorizing speech into separate subspaces (content, prosody, acoustic details, and timbre), FACodec allows to manage each attribute individually, leading to high-quality and adaptable speech synthesis. Here is an explanation of how each component in the diagram contributes to FACodec architecture and training process.

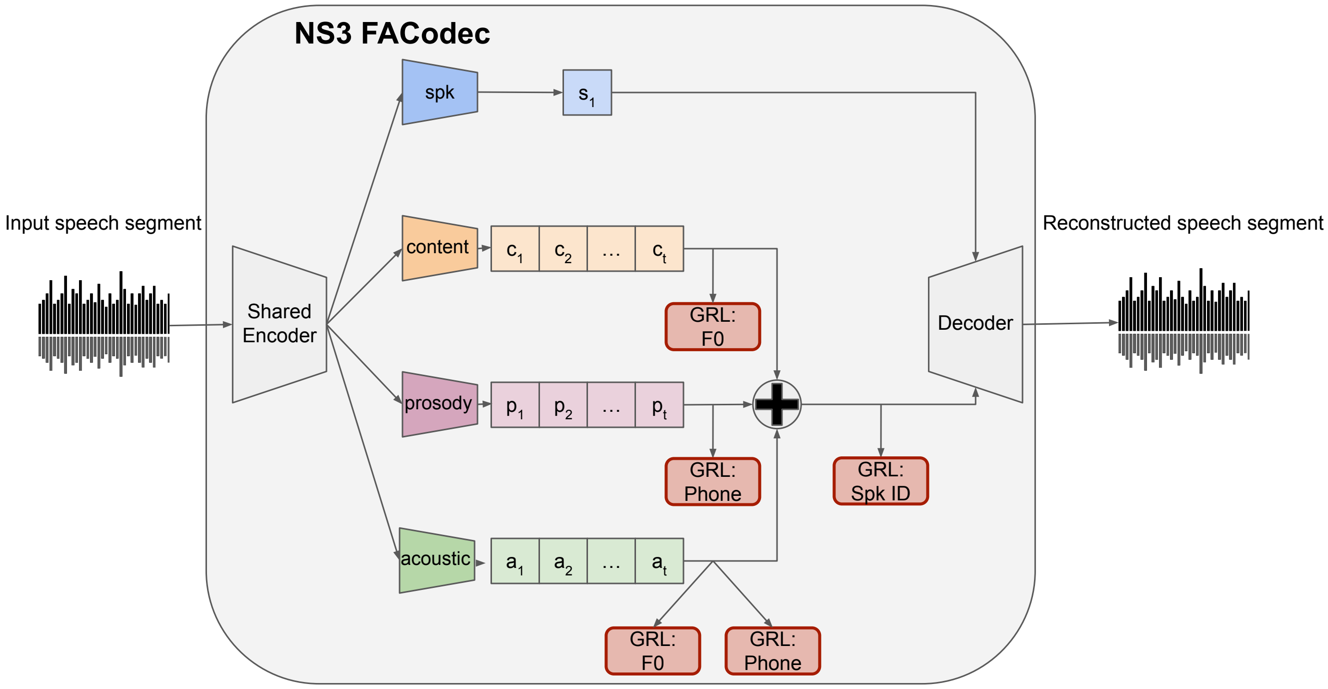


Figure 3.4: NaturalSpeech3 FACodec training pipeline.

- **Shared Encoder:** The Shared Encoder processes the input speech segment, capturing the initial latent representation of the audio. This encoder uses convolutional blocks with downsampling to reduce the complexity of the input and produce a dense, low-dimensional representation, effectively creating a bottleneck that limits the information flow and promotes disentanglement of attributes.

- **Speaker Encoder (spk):** The Timbre Encoder uses a Transformer-based module to create a global timbre vector ($s_1 \in \mathbb{R}^d_c$, where d_s is a speaker vectors' dimension) that captures speaker-specific characteristics.

- **Content Encoder (content):** The Content Encoder focuses on the linguistic content of the speech, encoding phonetic information as a sequence of tokens ($c_1, ..., c_t \in \mathbb{R}^d_c$, where d_c is a content vectors' dimension) via Factorized Vector Quantization (FVQ). Phoneme labels guide this encoder, ensuring it captures linguistic elements without prosodic information.

- **Prosody Encoder (prosody):** The Prosody Encoder captures the rhythm and intonation of speech, encoding these features as frame-level tokens ($p_1, ..., p_t \in \mathbb{R}^d_c$, where d_p is a prosody vectors' dimension) with FVQ. Supervised with pitch

(F0) targets, it captures pitch contours and rhythmic variations that define speech expressiveness.

- **Acoustic Detail Encoder (acoustic)**: The Acoustic Detail Encoder focuses on capturing subtle, low-level acoustic features ($a_1, ..., a_t \in \mathbb{R}^d_c$, where d_c is an acoustic vectors' dimension) that add realism to the synthesized speech. This encoder also uses FVQ for frame-level tokenization. Additionally, a detail dropout technique occasionally masks acoustic tokens during training, forcing the model to reconstruct intelligible speech without relying on acoustic details, which helps maintain attribute separation.
- **Decoder:** The Decoder receives the combined representations from content, prosody, acoustic details, and timbre to reconstruct the speech waveform. The integration of these separate attributes is achieved through conditional layer normalization on the timbre vector, ensuring that speaker identity is incorporated into the synthesized speech.

Next, we describe loss functions used for FACodec training. The diagram on Figure 3.4 highlights how Gradient Reversal Layers (GRLs) [113] play a key role in enforcing disentanglement between subspaces. These GRLs are adversarial layers that apply reverse gradients to specific subspaces, preventing certain types of information from leaking across attribute boundaries:

- **Content GRL (GRL: F0):** Blocks prosodic information in the content subspace.
- **Prosody GRL (GRL: Phone):** Blocks content information in the prosody subspace.
- **Acoustic GRL (GRL: F0 and Phone):** Blocks both prosody and content in the acoustic details subspace.
- **Combined GRL (GRL: Spk ID):** Blocks timbre information for all content, prosody and acoustic representations.

In addition to GRLs, FACodec employs supervised losses and various techniques to achieve effective disentanglement across the different speech attributes. **The information bottleneck mechanism** [115, 15] constrains each attribute representation to a low-dimensional space, which encourages the model to focus on essential information and promotes clearer separation between attributes. **Supervised losses** further reinforce this separation by using specific targets: phoneme labels guide the content subspace, pitch values guide prosody, and speaker ID targets ensure the timbre subspace captures speaker-specific characteristics. Additionally, **detail dropout** is applied during training, where the acoustic subspace is occasionally masked to ensure that the prosody, content, and timbre subspaces can independently reconstruct a recognizable speech segment. This dropout strengthens the disentanglement and prevents the model from over-relying on acoustic details alone.

3.1.5 Challenges in Disentanglement Learning for Speech.

While disentanglement has demonstrated significant potential in speech processing, several challenges remain. The primary challenge is the complexity of the speech signal itself, which is highly multi-dimensional, containing information about the speaker's identity, content, prosody, and emotion all at once. Isolating these factors reliably can be difficult, especially in noisy or real-world environments where overlapping information can confound models.

Another challenge is ensuring that disentangled representations are interpretable and generalizable. For example, models that disentangle speech for anonymization must ensure that the representations they produce are not only effective at anonymizing the speaker but also generalizable to different speakers, accents, and speech styles. Overcoming these challenges requires more robust models and novel training techniques that can balance the trade-offs between privacy, intelligibility, and accuracy.

3.2 Disentanglement approaches for VPC2024

To address the four privacy conditions (EER$_1$, EER$_2$, EER$_3$, EER$_4$) in the VPC2024, we enhance three distinct systems instead of relying on a single solution. This approach ensures a diverse set of solutions adjusted to different use cases. The three independent systems are based on NaturalSpeech3 FACodec, B3, and B5, respectively. The modifications applied to these systems, along with the corresponding results, are detailed in this section.

3.2.1 Baseline Systems

In this subsection, we summarize two baseline systems, B3 and B5, that we use in our experiments. A modified version of B3 is used to create submission for the first condition (EER$_1$) and the second condition (EER$_2$), while a modified version of B5 is used for the third condition (EER$_3$) and the last condition (EER$_4$). The details about all proposed changes are laid out in Subsection 3.2.2.

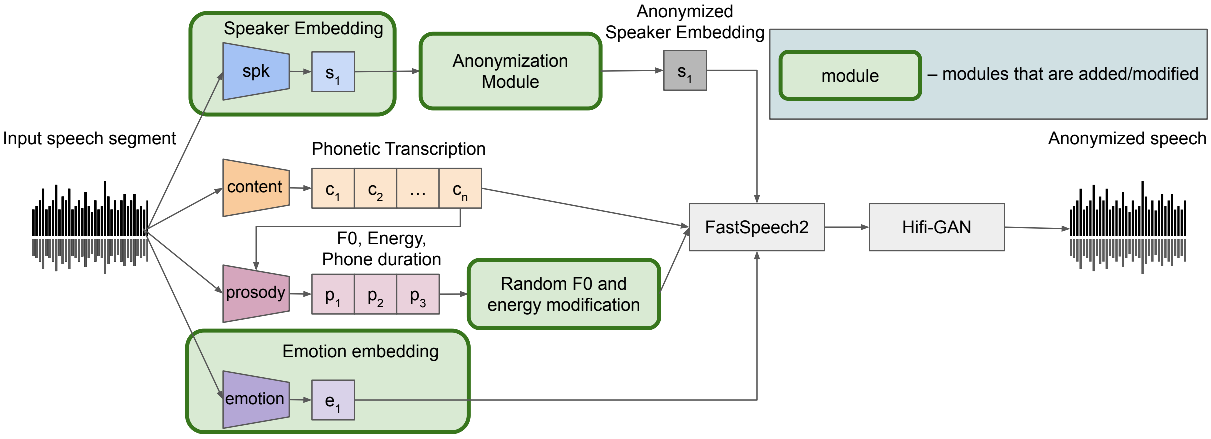


Figure 3.5: Schematic diagram of Modified system B3. Differences between baseline and modified systems are highlighted in green boxes.

B3. The baseline system B3 (Figure 2.10) uses a Wasserstein generative adversarial network with Quadratic Transport Cost (WGAN-QC) [96] to generate artificial pseudo-speaker embeddings, anonymizing the speaker's identity through four main steps:

- **Phonetic Transcriptions Extraction (content):** Phonetic transcriptions $(c_1, c_2, ..., c_n)$ are extracted using an end-to-end automatic speech recognition (ASR) model with a hybrid CTC-attention architecture.

- **Speaker Embeddings (spk):** Speaker embeddings (s_1) are obtained using an adapted Global Style Tokens (GST) model [93].

- **Prosody Extraction (prosody):** Extracted prosody consist of 3 components: phone durations, F_0 and energy. Phone durations are extracted using 6-ayer CNN and LSTM with CTC loss. F_0 is estimated using Praat.

- **Anonymization Module:** The original speaker embedding (s_1) is swapped with an artificial one generated by a WGAN. If the cosine distance between the artificial and original embeddings is less than 0.3, the replacement is considered successful. Otherwise, the process is repeated up to 30 times. Additionally, the pitch and energy values for each phoneme are adjusted using random values between 0.6 and 1.4.

- **Speech Synthesis:** The anonymized speaker embedding, adjusted prosody, and original phonetic transcription are used to create anonymized speech using the FastSpeech2 model and HiFi-GAN [90] vocoder, as implemented in IMS-Toucan [116].

B5. The B5 system used a HiFi-GAN model conditioned on fundamental frequency and a linguistic representation of the source utterance along with speaker embedding of a designated speaker to generate anonymized speech. The diagram of B5 is illustrated on Figure 2.12 above.

- **Fundamental Frequency (F0):** B5 uses a pytorch implementation of YAAPT Pitch Tracking [117] to extract F0 from speech. In authors' of B5 thesis [39], the original authors of B5 suggest several complementary normalization or transformations to be applied to F0, none of which are included in B5.

- **Linguistic Representation:** B5 uses the output of a vector quantization bottleneck layer (VQ-BN) put the top of the acoustic model (AM) of an automatic speech recognition (ASR) trained to classify left-biphone as the linguistic representation.

- **Speaker Embedding:** a designed speaker embedding of a speaker included in the training stage is used to change the voice of anonymized speech. We randomly pick a speaker embedding to anonymize each utterance similar to the B5 baseline provided by the organizer.

We use the same pre-trained B5 model provided by organizers without doing any further training or tuning. Instead, we introduce a new method of transformation that can be applied to F0 in the inference stage, which is discussed in Subsection 3.2.2

3.2.2 Our Methods

In this subsection, we elaborate on the details of proposed modifications to the baseline systems.

Modifications of B3.

The modified system B3 is illustrated in Figure 3.5. We experimented with the following main modifications for this system is as follows:

- Incorporated emotion embeddings via a fine-tuned Wav2Vec2 model as input to the FastSpeech2 model.

- The Global Style Tokens (GST) model is replaced by different speaker embedders such as WavLM [118] and ECAPA2 [119].

- Explored anonymization strategies like random speaker selection and cross-gender anonymization.

- Investigated the impact of varying prosody modification ranges on privacy and utility metrics.

We start off with emotion embeddings. For extracting emotion-based embeddings, we employ a fine-tuned Wav2Vec2 Large Robust model [120] on MSP-Podcast [121]. Notably, the model is pruned from 24 to 12 transformers, and the CNN component is frozen prior to fine-tuning. Embeddings are extracted from the hidden layer, which has 1024-dimensional vectors as output. We employ it to FastSpeech2 in the same way as speaker embeddings in [122] by adding one more linear projection and concatenating it with an output from Conformer [123].

In addition to the GST model, we implement different speaker embedding models such as ECAPA2 and WavLM with 128 and 512 embedding sizes correspondingly. As both these speaker embedders work on audios instead of spectrograms, we add a pre-trained HiFi-GAN to the setup for the second phase of FastSpeech2 training in order to generate audio and extract embeddings for cycle consistency loss. In contrast to the GST model, ECAPA2 and WavLM speaker embedders are frozen during FastSpeech2 training. Similarly, the HiFi-GAN is also frozen.

Furthermore, we explore various anonymization strategies, such as random speaker selection, which involves replacing the source speaker embedding for each utterance with a randomly selected embedding from a pool of embeddings. Additionally, we evaluated the importance of the usage of cross-gender for anonymization for the modified model. In cross-gender anonymization technique, we select a target speaker from the pool that has the opposite gender with respect to the source speaker. Finally, we examine how different powers of prosody anonymization affect privacy-based and utility-based metrics. To be more specific, we experiment with different offsets for pitch and energy multipliers.

The training process is the same as for the B3 baseline system. We retrain each part of the system except HiFi-GAN and ASR model. Detailed descriptions of the components are provided in Table 3.1. This table includes only the submitted models; models with different speaker embedding extractors were excluded because they were not selected as our final models. Further details on the experimental results can be found in Section 3.2.3.

Modifications of B5.

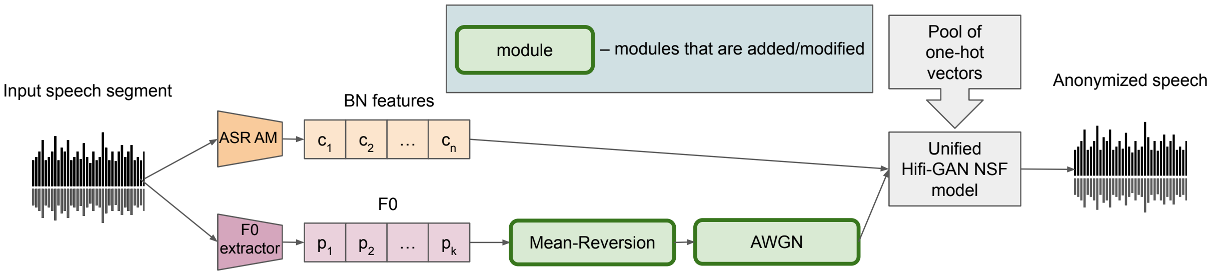


Figure 3.6: Schematic diagram of Modified system B5. Differences between baseline and modified systems are highlighted in green boxes.

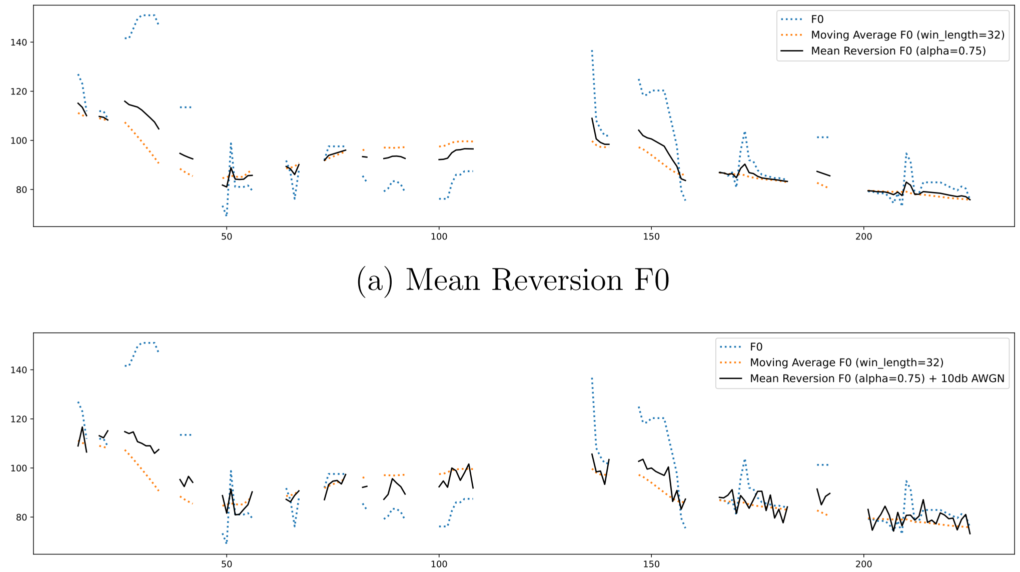
Table 3.1: Description of Systems 1b, 2a, 2b. The modules with differences between the baseline B3 system and modified B3 system are highlighted in green.

#	Module	Description	Output features	Data
1	Prosody extractor	Phone aligner: 6-layer CNN + LSTM with CTC loss F0 estimation using Praat F0, energy, durations normalized by each vector's mean	$F0^1$, $energy^1$ phone $durations^1$	LibriTTS: train-clean-100
2	ASR	End-to-end with hybrid CTC-attention Input: log mel $Fbank^{80}$ Encoder: Branchformer Decoder: Transformer CTC and attention criteria	phonetic transcript with pauses and punctuation	LibriTTS: train-clean-100 train-other-500
3	Speaker embedding extractor	GST, trained jointly with SS model Input: mel $spectrogram^{80}$ 6 hidden layers + 4-head attention	GST speaker $embedding^{128}$	LibriTTS: train-clean-100
4	Emotion embedding extractor	**1b, 2a:** Dimensional Speech Emotion Recognition Model based on Wav2vec 2.0 Input: Wav2vec 2.0 Large features	emotion $embedding^{1024}$	MSP-Podcast (v1.7)
		2b: –	–	–
5	Prosody modification module	**1b, 2b:** –	–	–
		2a: Value-wise multiplication of F0 and energy with random values in [0.7, 1.3)	$F0^1$, $energy^1$	LibriTTS: train-clean-100
6	Speaker anonymization module	**1b:** Averaged 100 embeddings randomly selected from a pool of 200 farthest embeddings from source by cosine scoring + cross-gender **2a, 2b:** Random Speaker selection per each source utterance + cross-gender	Anonymized speaker $embedding^{128}$	LibriTTS: train-clean-100
7	SS model	IMS Toucan implementation of FastSpeech2 Input: $F0^1$ + $energy^1$ + phone $durations^1$ + phonetic transcript + GST $embeddings^{128}$ (**1b, 2a:** + emotion $embeddings^{1024}$) Training criterion defined in FastSpeech2	mel $spectrogram^{80}$	LibriTTS: train-clean-100
8	Vocoder	HiFi-GAN vocoder Input: mel $spectrogram^{80}$ Training criterion defined in HiFi-GAN	speech waveform	LibriTTS: train-clean-100

For conditions EER_3 and EER_4, we base our submissions on the B5 [39] system provided by the organizer. Champion proposes a voice anonymization model using a HiFi-GAN vocoder which takes in the F0 and the linguistic representation of a source utterance and then turns it into the speech with the voice of a target speaker. In his thesis, the author explores different transformation techniques that can be applied to F0 including linear transformation, Additive White Gaussian Noise, and quantization. In this work, we propose a new type of transformation that uses the original and the n-frame moving average F0 ($\overline{F_0}$), with n=32 in our calculation:

$$\hat{F_0} = (1-\alpha)F_0 + \alpha\overline{F_0} \tag{3.2}$$

with $\alpha = 0$, we get the original F0, and with $\alpha = 1$, we get the moving average F0. For any α between 0 and 1, we obtain a mean reversion F0 which is a value weighted toward the mean. The motivation for this method is that we can reduce the dynamic range of F0, which is one characteristic of voice, and move it to a more neutral value. Moreover, the calculation is based on a short window instead of the entire utterance like the linear transformation method. Note that, we remove unvoiced frames when calculating the moving average F0.



(a) Mean Reversion F0

(b) Mean Reversion F0 with a 10-dB white gaussian noise

Figure 3.7: Examples of Mean Reversion F0 with and without additive noise

We then apply an Additive White Gaussian Noise on top of mean reversion F0 to push up the EER. Figure 3.7 shows an example of the mean reversion F0 with and without additive noise. Detailed information on the components is described in Table 3.2 and the diagram is illustrated on Figure 3.6.

Disentanglement-based models.

Table 3.2: Description of Systems 3 and 4. The modules with

#	Module	Description	Output features	Data
1	F0 extractor	F0 extracted with s pytorch implementation of YAAPT **3:** Using Mean Reversion F0 ($\alpha = 0.75$) in inference **4:** Using Mean Reversion F0 ($\alpha = 0.75$) and 10-db AWGN	F0	N/A
2	ASR AM with VQ	Acoustic Model trained to identify left biphones and a VQ bottleneck layer	Linguistic representation	VoxPopuli Librispeech: train-clean-100
3	Speaker embedding	One-hot vector represented speaker in training set	Speaker embedding	LibriTTS: train-clean-100
4	Speech Synthesis	HiFi-GAN vocoder Input: F0 + linguistic representation + speaker embedding	Speech waveform	LibriTTS: train-clean-100

Next, we explore disentanglement-based models, which are potentially valuable for removing speaker-related information from other components, such as prosody, content, and acoustic information. We compare two approaches: the β-VAE model and the NaturalSpeech3 (NS3) FaCodec. As shown in the diagram on Figure 3.4, NS3 FACodec achieves anonymization by first encoding the input speech segment into separate attributes—content, prosody, and acoustic details—using a shared encoder. Instead of retaining the original speaker embedding, the model replaces it with an anonymized speaker embedding (Spk Anon), which is then fed into the Modified HiFi-GAN synthesizer along with the other disentangled components. This process results in anonymized speech that preserves the linguistic and prosodic details of the input while effectively concealing the speaker's identity, thus enabling secure voice conversion. The detailed information of components is described in Table 3.3.

3.2.3 Experiments

In this section, we provide experimental results for the system modifications explained in Section 3.2.2. To begin, we present the results for different modifications of system B3.

Modified B3.

From the results in 3.4, it can be observed that emotion embeddings help to improve Emotion Recognition performance while maintaining ASR performance at the same level. However, there is some degradation in privacy, which might be due to speaker identity

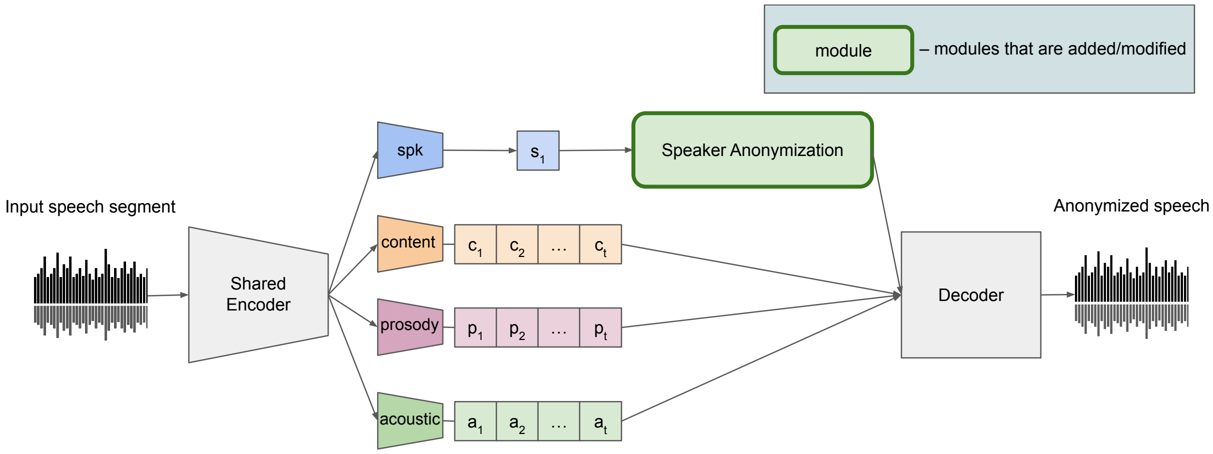


Figure 3.8: Schematic diagram of adapted NS3 FACodec for Anonymization. Differences between baseline and modified systems are highlighted in green boxes.

Table 3.3: Description of Systems 1a. The modules with differences between the baseline NaturalSpeech3 FACodec system and modified system are highlighted in green.

#	Module	Description	Output features	Data
1	Encoder [124]	4 Downsampling Convolution-based Layers with Snake activation function Input: speech waveform	Output vector^{256}	Librilight train [125]
2	Speaker embedding extractor	Several Conformer blocks	Speaker embedding^{256}	Librilight train
3	Content extractor	Factorized Vector Quantization (FVQ) with 2 quantizers, codebook size: 1024	Content vector^{256}	Librilight train
4	Prosody extractor	FVQ with 1 quantizer, codebook size: 1024	Prosody vector^{256}	Librilight train
5	Acoustic extractor	FVQ (FVQ) with 3 quantizer, codebook size: 1024	Acoustic vector^{256}	Librilight train
6	Speaker anonymization module	Averaged 100 embeddings randomly selected from a pool of 200 farthest embeddings from source by cosine scoring AWGN with scale= 0.075 Cross-gender	Anonymized speaker embedding^{256}	LibriTTS: train-clean-100
7	Decoder [124]	Upsampling Convolution-based Layers with Snake activation function	speech waveform	Librilight train

leakage in the emotion embeddings. In addition, we provide experimental results for this system without prosody anonymization to check how modifications of prosody affect SER performance. As shown in the results, removing prosody modifications improves SER and ASR but also reduces privacy, making this system suitable for condition with a minimum $\mathrm{EER}_1 = 10\%$.

Table 3.5 shows a comparison between WGAN and Random-Speaker anonymization

Table 3.4: Comparison between systems with and without emotion embedding with different speaker embedder and prosody anonymization. + in speaker anonymization column corresponds to Random-Speaker selection from LibriTTS-clean-100 for each source utterance. + in prosody anonymization column corresponds to the systems with prosody multipliers from $[0.6, 1.4]$ range.

Speaker Anon	Speaker Embed	Prosody Anon	Emotion Embed	EER		UAR		WER	
				dev	test	dev	test	dev	test
−	−	−	−	5.72	4.59	69.08	71.06	1.80	1.85
+	GST	+	−	25.76	28.42	37.97	37.39	4.33	4.33
+	GST	+	+	22.59	24.09	42.52	41.74	4.39	4.40
+	GST	−	+	16.88	17.45	42.76	43.21	**3.81**	**3.83**
+	WavLM	−	+	17.97	16.64	**43.84**	**45.67**	4.54	4.69
+	ECAPA2	−	+	19.48	22.55	42.53	42.37	4.83	4.80

Table 3.5: Comparison between WGAN anonymization strategy trained on LibriTTS-clean-100 and Random Speaker (Rnd-Spk) selection from LibriTTS-train-clean-100 per each source utterance.

Anon Type	EER		UAR		WER	
	dev	test	dev	test	dev	test
WGAN	25.20	27.78	38.40	37.70	4.30	4.40
Rnd-Spk	**25.76**	**28.42**	37.97	37.39	4.33	4.33

techniques. There is almost no difference in the privacy and utility metrics for these methods, so we chose to stick with Random-Speaker as it requires no training.

Next, Table 3.6 compares different ranges for multipliers involved in prosody anonymization. The results indicate that fewer prosody modifications result in worse privacy but better utility. This finding is useful for VPC2024 as it allows us to find a better trade-off between privacy and utility for specific EER conditions.

Modified B5.

Table 3.7 lists the results of the Mean Reversion F0 method discussed in Section 3.2.2 given different α values. We can see a general trend that EER increases when α is

Table 3.6: Comparison between difference range for F0 and energy multipliers. The bottom row corresponds to the system without prosody manipulation.

Multiplier Range	EER		UAR		WER	
	dev	test	dev	test	dev	test
$[0.6, 1.4]$	25.76	28.42	37.97	37.39	4.33	4.33
$[0.7, 1.3]$	23.93	25.62	37.49	37.59	4.07	4.05
$[0.8, 1.2]$	22.70	25.92	38.01	37.96	3.89	3.91
$[0.9, 1.1]$	19.88	22.62	39.03	37.17	3.80	3.77
−	19.47	21.82	38.91	38.11	3.70	3.75

Table 3.7: Evaluation results of B5 using Mean Reversion F0 with different values of α in inference stage

α	EER		UAR		WER	
	dev	test	dev	test	dev	test
0.00	31.64	31.36	39.18	38.24	4.79	4.44
0.25	32.13	32.03	**39.61**	38.38	4.74	4.54
0.50	33.48	34.08	38.60	37.34	**4.62**	4.54
0.75	**38.56**	37.48	38.06	37.60	4.70	4.47
1.00	37.91	**37.93**	38.50	**38.78**	4.79	**4.43**

Table 3.8: Evaluation results of B5 using Mean Reversion F0 ($\alpha = 0.75$) and AWGN with different magnitude of noise in inference stage

dB	EER		UAR		WER	
	dev	test	dev	test	dev	test
0	38.56	37.48	38.06	37.60	4.70	4.47
5	39.58	40.00	38.91	37.12	4.67	4.49
10	42.46	**43.15**	**39.41**	**38.47**	**4.63**	**4.40**
15	**42.97**	40.36	38.50	37.49	4.66	4.50
30	41.43	39.62	38.41	37.88	4.77	4.64

increased while UAR and WER fluctuate but not very significantly. We submitted the sample generated with $\alpha = 0.75$ for the condition EER_3.

For the last condition EER_4, we add a 10-db AWGN to the mean reversion F0 with $\alpha = 0.75$ and manage to obtain an EER above 40%. The result can be found in Table 3.11.

We note that the EER results of these systems were highly volatile during our experiments, often producing different results even if we ran with the same configuration. It seems that convergence of an attacker ASV model depends on factors such as the machine, GPU, randomly picked speaker embedding, and other random parameters. The systems that we selected for submissions were based on the results available at that time and represented our methods of Mean Reversion F0 and AWGN.

Disentanglement-based models

The comparison results between ß-VAE and NaturalSpeech3 FACodec are shown in Table 3.9. It can be seen that ß-VAE performs poorly in utility-based tasks, likely because of the fact that content representations are not rich enough.

As one might notice from the results in Table 3.9, NaturalSpeech3 has decent utility results. Therefore, we decided to employ anonymization techniques to improve privacy protection for NS3, aiming to meet a condition with minimum $\text{EER}_1 = 10\%$. We experimented with the following tricks: Additive White Gaussian Noise (AWGN) to Speaker Embedding and conversing a source speaker to a target speaker of the opposite gender

Table 3.9: Comparison between ß-VAE and NS3 disentanglement models.

Model	EER		UAR		WER	
	dev	test	dev	test	dev	test
Original	5.72	4.59	69.08	71.06	1.80	1.85
ß-VAE	10.71	10.49	30.38	31.28	67.72	65.5
NS3	9.29	8.78	51.64	52.89	2.97	2.77

Table 3.10: Comparison between NaturalSpeech3 FACodec systems with different power of AWGN applied to speaker embedding. The speaker anonymization module consists of averaging 100 embeddings randomly selected from a pool of 200 farthest embeddings (LibriTTS-train-clean-100) from source utterance by cosine scoring.

scale,	Cross	EER		UAR		WER	
$10e^{-3}$	Gender	dev	test	dev	test	dev	test
–	–	9.29	8.78	51.64	52.89	2.97	2.77
75	–	12.25	9.14	48.00	48.09	4.66	4.63
75	+	12.09	10.46	49.20	49.12	4.97	4.60
78	+	12.42	10.24	49.10	49.39	5.44	5.07
80	–	12.63	9.42	47.33	48.35	5.37	5.40
80	+	13.66	10.10	48.82	48.95	5.69	5.37
90	–	12.41	10.45	47.61	47.10	7.04	6.45

(cross-gender). The results are shown in Table 3.10. As we can see, cross-gender conversion helps to improve privacy and ASR performance on the corresponding test sets. Interestingly, it also improves SER performance on both development and test sets. As expected, AWGN enhances privacy at the cost of utility.

Our results underscore the need for a balance between privacy and utility, as methods like AWGN and prosody anonymization can strengthen privacy but also impact system performance. This balance is essential for creating anonymization techniques that are both secure and functional.

Submitted systems

In this subsection, we provide a summary of all submitted systems. Table 3.11 shows privacy and utility results for each of the conditions.

Additionally, we prepared the tables 3.3, 3.1, 3.2 with a summary of architecture, input, output values, and training data for components in submitted systems. Furthermore, we provide the comparison between baseline and submitted systems in Figures 3.10 and 3.9.

Table 3.11: Results summary for all submitted systems grouped by achieved privacy conditions.

Condition	System	EER		UAR		WER	
	ID	dev	test	dev	test	dev	test
EER_1	1a	12.09	10.46	49.20	49.12	4.97	4.60
	1b	16.88	17.45	42.76	43.21	3.81	3.83
EER_2	2a	21.47	24.13	44.67	42.78	4.21	4.29
	2b	20.07	22.85	39.18	37.67	3.61	3.68
EER_3	3	38.56	37.48	38.06	37.60	4.70	4.47
EER_4	4	42.46	43.15	39.41	38.47	4.63	4.40

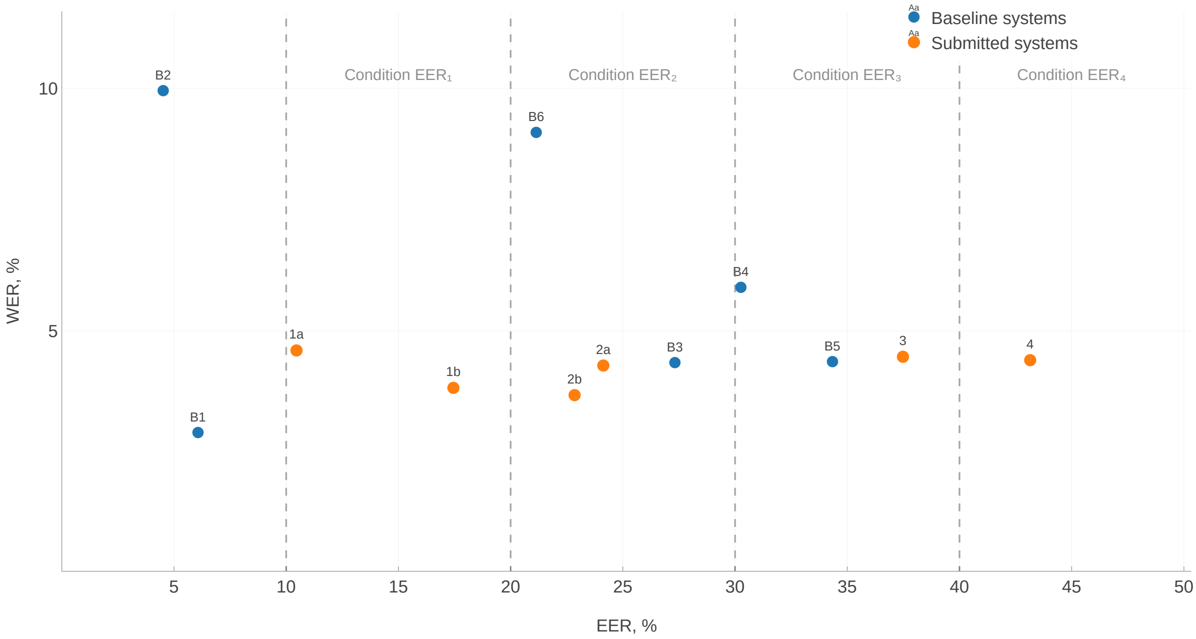


Figure 3.9: Comparison between baseline and submitted systems in terms of Privacy and ASR.

3.3 Conclusion

In this chapter, we examine how disentanglement learning for speech processing can separate linguistic and paralinguistic features, thereby enhancing anonymization. We provide an overview of approaches like VAEs and Neural Codecs, and demonstrate how they support tasks such as speaker verification, emotion recognition, and voice conversion. We then outline our contributions to the Voice Privacy Challenge 2024, where instead of introducing a single new anonymization system, we refine the provided baselines to satisfy all requirements and boost performance metrics. This involves adding emotion embeddings, testing WavLM and ECAPA2 speaker encoders for the B3 baseline, comparing speaker and prosody anonymization methods, and proposing a Mean Reversion F0 strategy for B5. We also investigate β-VAE and NaturalSpeech3 FACodec for improved

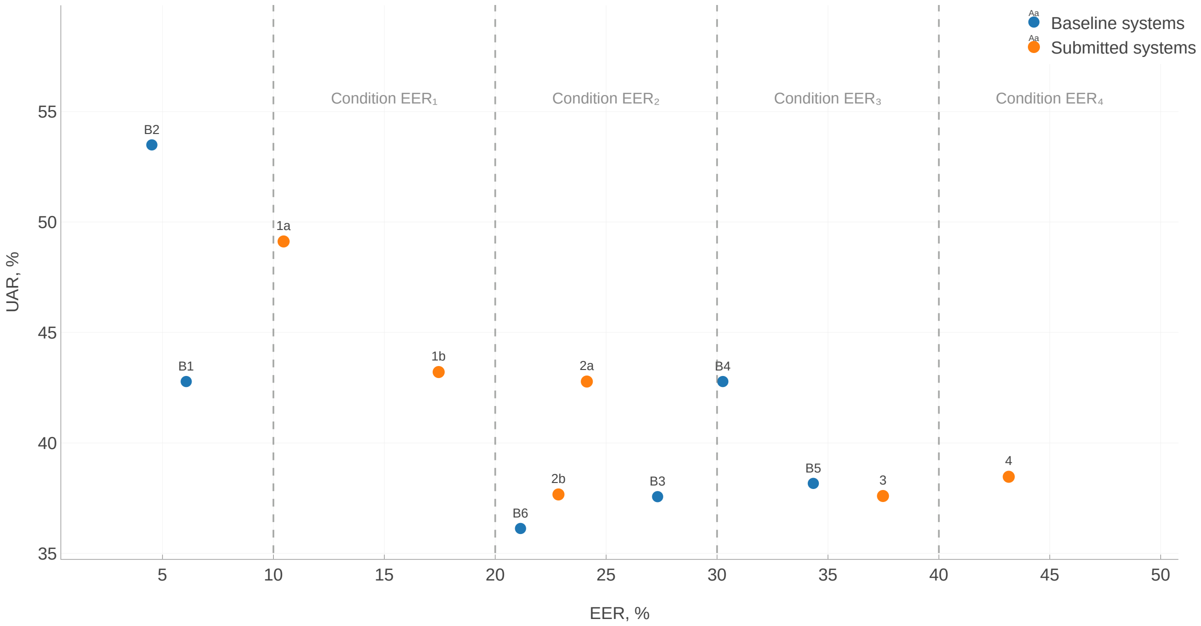


Figure 3.10: Comparison between baseline and submitted systems in terms of Privacy and SER.

disentanglement. By adjusting these approaches and baselines, we create submissions adjusted to each EER condition—leveraging a modified NaturalSpeech3 FACodec for EER_1, a refined B3 for EER_1 and EER_2, and an enhanced B5 for EER_3 and EER_4.

Chapter 4

Multi-Enrollment Speaker Recognition

4.1 Introduction

In this chapter, we consider a general scenario that we call *online speaker recognition*, where speech segments arrive sequentially, and the speaker recognition system has to identify previously encountered speakers and detect new speakers. At each time, there is a history of previously processed segments and the current segment to be classified.

One application scenario is *household speaker recognition* [126, 127]. A household is a small set of family members whose speech data is processed by a shared device such as a smart speaker (*e.g.* Amazon Alexa). First, the device collects speech data from the users to create their profiles (speaker models). Then, at each interaction with a person, the device identifies the user and, optionally, updates (enriches) the corresponding speaker model. The device continuously collects the data of the users to improve its performance by estimating more accurate speaker representations. Also, the recorded speech utterances may belong to unregistered speakers (*e.g.* guests) leading to an open-set identification task. Another related task is low-latency speaker spotting [128], where a previously registered target speaker has to be detected in an audio stream.

Another example is *online speaker diarization* or *clustering* [129, 130, 131, 132, 133, 134]. In this case, short speech segments from an audio stream have to be classified with low latency (*e.g.* 1-2 seconds). Unlike household speaker recognition, where all unregistered speakers are not of interest, in the speaker clustering task, there are no speakers registered beforehand, and a new speaker model has to be created for each previously unseen speaker. In the following, we focus on the online speaker clustering task since it is more general, and online speaker recognition can be seen as a special case.

What these scenarios have in common is that speech segments are received *sequentially* in nature and have to be classified on arrival. Specifically, an *open-set identification* problem has to be solved for each new segment. That is, the current segment has to be assigned to either one of the known speakers or a new (unknown) speaker. As a result, the number of segments per speaker continuously increases over time. This requires some way to aggregate information from multiple segments to form a memory-efficient speaker representation. This is usually referred to as *multi-enrollment* (or multi-session) speaker recognition [135, 136, 137, 138], that is, when a speaker is represented by multiple speech segments. Moreover, different speakers may be represented by *different* numbers of segments. As shown in [136], this can be a major complicating factor for speaker recognition, since it causes inconsistency in scores from different speaker models. To our best knowledge, this issue has not been studied for modern large-margin speaker embeddings.

Inspired by [136, 139], this work focuses on the issues arising from multi-enrollment scoring since it is a core element of online speaker recognition and clustering. We show that popular cosine scoring could have undesirable properties when used for multi-enrollment verification. Then we show that a highly constrained version of PLDA can be a suitable alternative while having better performance and comparable computational complexity. Specifically, we propose a PLDA model with spherical between- and within-covariance matrices as a replacement for cosine scoring back-end. While being *equivalent* to cosine scoring in a special case, this model can naturally handle varying degrees of uncertainty specific to the multi-enrollment scenario.

Further, we propose a probabilistic back-end for online speaker recognition and clustering. It is based on the spherical PLDA model and therefore has several appealing properties compared to cosine scoring. It employs an incremental (online) variant of variational Bayesian inference and provides probabilistic soft decisions for each input observation, based on the history of preceding observations.

Our contributions are summarized as follows:

- We compare scoring back-ends for multi-enrollment verification for modern large-margin embeddings.

- We propose a simple alternative to cosine scoring suitable for multi-enrollment verification.

- We propose a probabilistic back-end for online speaker recognition and clustering.

4.2 Background

4.2.1 PLDA

General formulation. In this study we focus on a variant of PLDA known as the *two-covariance model* [140]. Let $\mathbf{x}_{i,j} \in \mathbb{R}^d$ denote the jth speaker embedding of speaker i. Also, let $\mathbf{y}_i$ be the latent speaker identity of speaker i. Then, the model is specified by two Gaussian distributions:

$$p(\mathbf{y}_i) = \mathcal{N}(\mathbf{y}_i|\boldsymbol{\mu}, \mathbf{B}), \quad p(\mathbf{x}_{i,j}|\mathbf{y}_i) = \mathcal{N}(\mathbf{x}_{i,j}|\mathbf{y}_i, \mathbf{W}). \tag{4.1}$$

Here, $\boldsymbol{\mu}$ is a global mean, and $\mathbf{B}, \mathbf{W} \in \mathbb{R}^{d\times d}$ are the between- and within -speaker covariance matrices, respectively.

Being a linear Gaussian model, PLDA allows making inferences about speaker identities in closed form. Given a set of observations (embeddings), one can compare different hypotheses about the partition of this set by computing the corresponding hypothesis likelihoods. This is often referred to as *by-the-book* scoring in the literature [135, 141].

PLDA with spherical covariances. Despite being a gold standard for previously popular i-vectors [142], one could recently observe a gradual shift towards replacing PLDA with a simpler parameter-less cosine scoring back-end [143]. As discussed in [139], the high intra-speaker compactness of the large-margin embedding makes the conventional full-rank PLDA model superfluous. It was also observed in [139] that discarding off-diagonal elements in the within-speaker covariance matrix can bring considerable performance gain. Here, we analyze a much more constrained version of the PLDA model, to our knowledge, firstly proposed in [126, 144]. Specifically, we consider PLDA with *spherical covariances*, $\mathbf{B} = \sigma_{\mathrm{B}}^2\mathbf{I}$, $\mathbf{W} = \sigma_{\mathrm{W}}^2\mathbf{I}$, where σ_{B}^2 and σ_{W}^2 are between- and within-speaker variances and $\mathbf{I}$ denotes an identity matrix. In the following text, we will refer to this model as the *spherical PLDA*.

Relationship with cosine scoring. As was shown in [143], for length-normalized and centered embeddings, the verification likelihood ratio of the spherical PLDA can be written as a scaled and shifted cosine similarity measure. Since an affine transformation of scores is order-preserving, the two scoring rules are equivalent. This brings up a question about the usefulness of spherical PLDA. As we discuss further, spherical PLDA has several advantages over cosine scoring. For instance, we show that the PLDA by-the-book scoring outperforms different cosine based heuristic scoring methods in multi-enrollment verification.

Relationship with PSDA. Another closely related scoring back-end is the so-called probabilistic spherical discriminant analysis (PSDA) recently proposed in [145]. It can be viewed as PLDA model with Gaussian distributions replaced by von Mises-Fisher (VMF) distributions that are defined on the $d-1$ dimensional unit hypersphere $\mathbb{S}^{d-1}$ [146]:

$$p(\mathbf{y}_i) = \mathcal{V}(\mathbf{y}_i|\boldsymbol{\mu}, b), \quad p(\mathbf{x}_{i,j}|\mathbf{y}_i) = \mathcal{V}(\mathbf{x}_{i,j}|\mathbf{y}_i, w). \tag{4.2}$$

Here, $\mathcal{V}(\mathbf{y}|\boldsymbol{\mu}, \kappa)$ denotes the density of the VMF distribution with mean direction vector $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ and scalar concentration $\kappa \geq 0$ parameter. Similar to spherical PLDA, this model is parameterized by the mean direction vector $\boldsymbol{\mu}$ and two scalars: between-speaker, b, and within-speaker, w, concentrations.

The relation to spherical PLDA follows from the fact that restricting any isotropic Gaussian density to the unit hypersphere gives a VMF density, up to normalization. However, the two models are *not* equivalent, though their behavior is very similar as we show in the experiments.

We use both spherical PLDA and PSDA as a basis for a proposed online speaker clustering algorithm described in Section 4.3.

4.2.2 Multi-enrollment verification

When multiple enrollment utterances are available, they may represent different acoustic environments or channels, which can help disentangle speaker identity from other irrelevant factors. As shown in the diagram on Figure 4.1, the multi-enrollment speaker verification pipeline involves using multiple enrollment segments $(e_1, e_2, ..., e_K)$, each potentially carrying unique environmental and channel characteristics, to form a more robust speaker representation. For each test trial $(t_1, t_2, ..., t_N)$, speaker embeddings are extracted, and a Fusion process is applied to aggregate information from both test and enrollment embeddings, producing a similarity score. If only a single enrollment $(K = 1)$ is available, the system performs single-enrollment verification; if multiple enrollments $(K > 1)$ are used, the fusion leverages information from multiple sources, improving

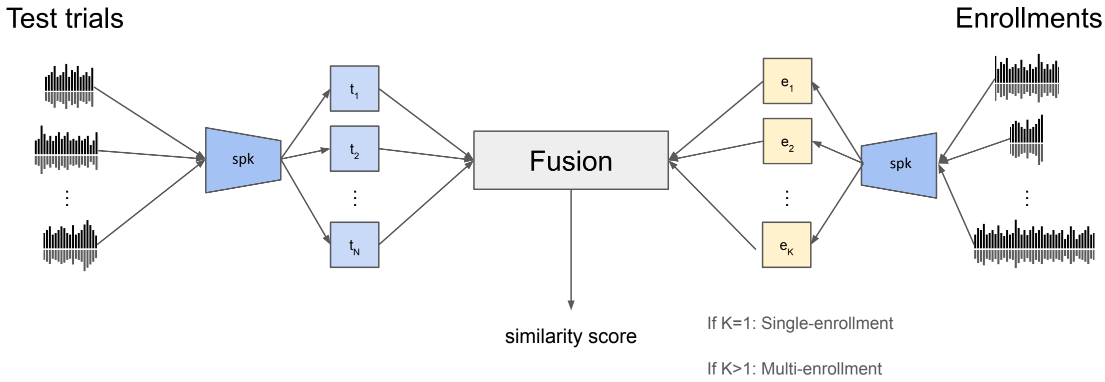


Figure 4.1: Evaluation pipeline of Multi-Enrollment Speaker Verification

verification accuracy. Prior studies, such as [135], compared various aggregation methods—including embedding averaging, score averaging, and by-the-book scoring—finding that embedding averaging performed best for i-vectors. However, our experiments reveal that for modern large-margin embeddings, by-the-book scoring with spherical PLDA or PSDA offers superior performance, suggesting a shift in optimal aggregation methods as embeddings evolve.

4.2.3 Evaluation Metrics

This subsection describes evaluation metrics that are used in our experiments.

minDCF.

First of all, we introduce additional metric to EER for Speaker Verification – minDCF. While Equal Error Rate (EER) is often easy to compute, it does not always suit practical applications as it treats false acceptance and false rejection errors equally, which may not align with real-world priorities. To address this, the Detection Cost Function (DCF) is used to incorporate the costs associated with both types of errors. DCF quantifies the weighted impact of the False Rejection Rate (FRR) and False Acceptance Rate (FAR) at a specific decision threshold, denoted as θ, and is expressed by the formula in Equation 4.3.

$$\mathrm{DCF} = C_{\mathrm{FR}} P(\mathrm{FR}|\mathrm{Tar}, \mathrm{Thres} = \theta) P_{\mathrm{Tar}} + C_{\mathrm{FA}} P(\mathrm{FA}|\mathrm{NonTar}, \mathrm{Thres} = \theta)(1 - P_{\mathrm{Tar}}), \quad (4.3)$$

where C_{FR} and C_{FA} are the respective costs for false rejection and false acceptance errors, $P(\mathrm{FR}|\mathrm{Tar}, \mathrm{Thres} = \theta)$ represents the probability of a false rejection given a target speaker, and $P(\mathrm{FA}|\mathrm{NonTar}, \mathrm{Thres} = \theta)$is the probability of a false acceptance given a non-target. The term P_{Tar} reflects the prior probability of encountering a target speaker. These parameters (C_{FR}, C_{FA} and P_{Tar}) are typically predefined and can be adjusted to reflect the relative importance of different error types, based on the specific application.

In practical scenarios, such as in voice-based authentication systems for banking, a higher emphasis might be placed on minimizing false acceptances to secure access to sensitive accounts. In this context, the cost associated with false acceptance, C_{FA}, can be increased to reflect its higher significance.

The minimum Detection Cost Function (minDCF) is commonly used as a performance metric to find the lowest achievable DCF by adjusting the decision threshold θ. This metric is calculated by varying the threshold across a range of values to identify the point where DCF is minimized. Unlike DCF at a fixed threshold, minDCF reflects the best possible trade-off between false acceptances and false rejections for a system, making it a useful benchmark for comparing models and optimizing system performance in speaker verification tasks.

DER and JER.

Secondly, we describe metrics for Online Clustering and Speaker Diarization: Diarization Error Rate (DER) and Jackard Error Rate (JER).

Diarization Error Rate (DER) is a commonly used metric in speaker diarization that measures the accuracy of speaker labeling in an audio recording. DER is calculated as the sum of three error components: Speaker Confusion (SC), where a speaker's identity is incorrectly assigned; False Alarm (FA), where non-speech regions are incorrectly labeled as speech; and Missed Speech (MS), where actual speech is marked as non-speech. These components are summed and divided by the total duration of the audio segments, as shown in the formula in Figure 4.2.

In the diagram, DER is visually broken down by segments. The Reference row shows the actual speakers, while the Hypothesis row represents the diarization model's output. Misalignments between these rows indicate various errors: SC occurs when speaker labels are swapped (e.g., speaker 1 labeled as speaker 2), FA appears when non-speech is marked as speech, and MS is observed where actual speech goes undetected. A DER of 0% indicates perfect performance, while higher values denote more errors.

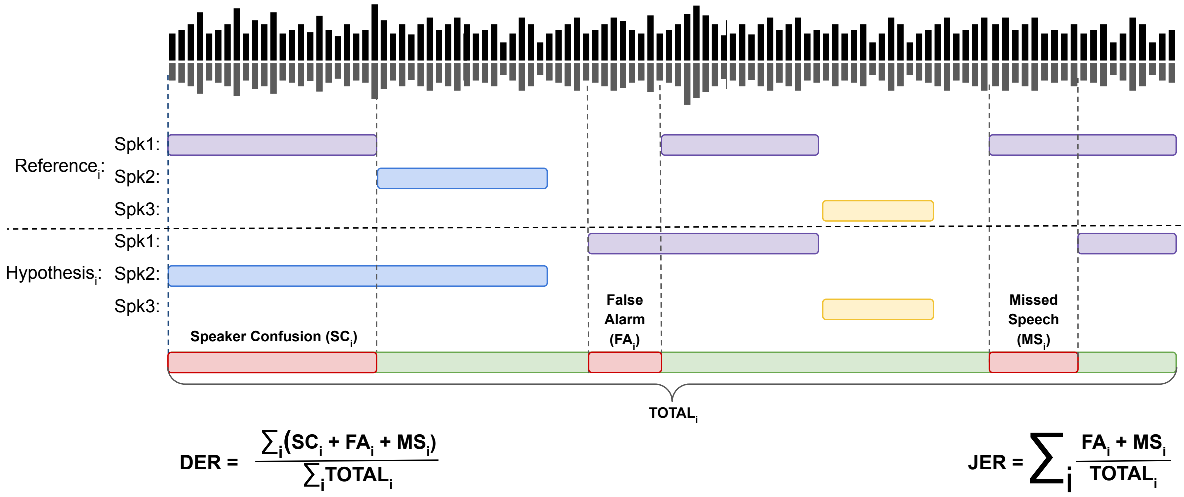

Spk1:
Reference$_i$:
Spk2:
Spk3:
Spk1:
Hypothesis$_i$:
Spk2:
Spk3:
Speaker Confusion (SC$_i$)
False Alarm (FA$_i$)
Missed Speech (MS$_i$)
TOTAL$_i$
DER = $\frac{\sum_i(SC_i + FA_i + MS_i)}{\sum_i TOTAL_i}$
JER = $\sum_i \frac{FA_i + MS_i}{TOTAL_i}$

Figure 4.2: DER and JER scoring.

Jaccard Error Rate (JER), on the other hand, is a metric based on the Jaccard index and was introduced in DIHARD II for assessing diarization accuracy. It measures the overlap between reference and system speaker segments, focusing on the union and intersection of their durations. JER is calculated by pairing each reference speaker with a system speaker using an optimal mapping and computing the ratio of missed and false alarm durations to the total duration for each speaker pair.

In the diagram, JER scoring is illustrated by comparing overlap between segments for each speaker in the reference and hypothesis rows. It captures both FA and MS errors but does not account for speaker confusion, making it particularly useful for cases where speaker detection accuracy is prioritized over identity accuracy. JER values range from 0% (perfect overlap) to 100% (no overlap), and it often correlates with DER, though it may provide additional insights in cases with significant speaker overlap or missed detections.

4.3 Online probabilistic speaker clustering

In this section, we describe the proposed back-end model for online speaker recognition and clustering. The difference between offline (batch) and online settings is that in the former case all the data to be processed is available at once, while in the latter case pieces of data are observed sequentially, in some order.

4.3.1 Online clustering

The general pattern behind many online clustering algorithms is solving a series of successive open-set identification tasks [129, 128, 131, 147]. The basic idea is to compare each new observation to the existing clusters, and either alter the closest cluster or create a new cluster. The generic Algorithm 1 demonstrates this for a single time step t.

Algorithm 1 Online clustering (time step t)

```
$s_i \leftarrow \text{score}(\mathbf{x}_t, \mathbf{X}_i)$        ▷ Compare $\mathbf{x}_t$ to the existing clusters
$k \leftarrow \arg\max_i s_i$        ▷ Find the most similar cluster
if $s_k \geq \tau$ then        ▷ If the maximal score $s_k$ is above the threshold $\tau$
    $\mathbf{X}_k \leftarrow \{\mathbf{X}_k, \mathbf{x}_t\}$        ▷ Add $\mathbf{x}_t$ to the $k$-th cluster
else
    $\mathbf{X}_{K+1} \leftarrow \{\mathbf{x}_t\}$        ▷ Create a new cluster
    $K \leftarrow K + 1$        ▷ Increment the total number of clusters
end if
```

First, the observation $\mathbf{x}_t$ is compared to all existing clusters, each represented by a set of observations, $\mathbf{X}_i$. If similarity to the closest cluster is above the threshold, τ, then $\mathbf{x}_t$ is assigned to this cluster. Otherwise, a new cluster is formed.

In this algorithm, clusters are represented by subsets of observations sharing the same label. Therefore, computing similarity to a cluster involves many-to-one comparison, also referred to as multi-enrollment verification in the context of speaker recognition.

As discussed in [136], varying cluster sizes may result in miscalibrated scores leading to sub-optimal decisions with a fixed threshold τ.

We aim at addressing this issue and propose an algorithm suitable for online clustering. Specifically, the underlying scoring model should be robust to varying cluster sizes naturally occurring in the online scenario. The proposed algorithm can be seen as a probabilistic extension of the Algorithm 1 constructed upon PLDA or PSDA models. As a result, it benefits from the advantages of PLDA (or PSDA) for multi-enrollment verification.

4.3.2 Model-based clustering

We start with a brief description of a generative model-based clustering [148, 149].

Model-based clustering builds upon a generative model that specifies how a set of data points $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ is generated from the hidden parameters of K clusters $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_K\}$, given the cluster assignments $\mathbf{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\}$. A typical clustering model is given by the following joint distribution: $p(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = p(\mathbf{X}|\mathbf{Z}, \mathbf{Y})p(\mathbf{Y})p(\mathbf{Z})$.

The clustering problem requires finding the most likely partition of the data $\mathbf{Z}_* = \arg\max_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X})$. Our approach is based on the "mean-field" variational Bayesian approximation [150, 151] assuming that the approximate posterior factorizes as $p(\mathbf{Z}, \mathbf{Y}|\mathbf{X}) \approx q(\mathbf{Z})q(\mathbf{Y})$. This assumption leads the algorithm consisting of iterative updates of the factors $q(\mathbf{Z})$ and $q(\mathbf{Y})$.

However, such updates are designed for the conventional clustering setup, where all observations are available *at once*. We modify the standard inference algorithm to make it suitable for *online* clustering, where observations arrive sequentially. This algorithm can be seen as an online version of the VBx [152] with simplified prior on assignments $p(\mathbf{Z})$. It is also similar to the algorithm from [132], where the authors modified the offline variational inference to make it suitable for online processing.

4.3.3 The proposed algorithm

Let us denote the current observation at the time step t as $\mathbf{x}_t$, and use the notation $\mathbf{X}_{1:t} = \{\mathbf{x}_1, ..., \mathbf{x}_t\}$ to denote causal observations.

The algorithm updates posterior distributions of latent identity variables $q(\mathbf{y}_k) \approx p(\mathbf{y}_k|\mathbf{X}_{1:t})$ after receiving a new observation $\mathbf{x}_t$. In general, several update iterations can be done. Our experiments reveal that even a single update can be sufficient for reasonable performance. In this case only posterior for the current data point $q(\mathbf{z}_t)$ needs to be computed, followed by updating each of $q(\mathbf{y}_k)$:

$$q(\mathbf{z}_t) \propto \exp \sum_{k=1}^{K} z_{t,k} \underbrace{\left[\mathbb{E}_{q(\mathbf{y}_k)}[\log p(\mathbf{x}_t|\mathbf{y}_k)] + \log \pi_k\right]}_{\log \gamma_{t,k}}, \tag{4.4}$$

$$q(\mathbf{y}_k) \propto \exp\left[\gamma_{t,k} \log p(\mathbf{x}_t|\mathbf{y}_k) + \log q(\mathbf{y}_k|\mathbf{X}_{1:t-1})\right]. \tag{4.5}$$

Here, $\gamma_{t,k}$ is the k-th component of the vector of posterior probabilities $q(\mathbf{z}_t)$ over the cluster assignments and π_k are the corresponding prior probabilities.

This algorithm continuously updates speaker models defined by $q(\mathbf{y}_k)$. Also, one can obtain speaker labels at each time step t by finding $\arg\max_k \gamma_{t,k}$. For instance, if $\gamma_{t,k} = 0$, then the posterior $q(\mathbf{y}_k)$ stays unchanged. In general, if the soft-assignments $\boldsymbol{\gamma}_t$ were converted into hard decisions, then updating $q(\mathbf{y}_k)$ would be nothing more than the sequential application of the Bayes formula. Also, the algorithm would become very similar to a sequence of multi-enrollment recognition tasks, where predictions are obtained via by-the-book scoring.

These update equations can be used to construct different online recognition and clustering algorithms depending on a particular choice of the underlying generative model defined by $p(\mathbf{x}|\mathbf{y})$ and $p(\mathbf{y})$. In this study, we use two models: spherical PLDA and PSDA. Table 4.1 demonstrates the update equations for both models.

PLDA: $q(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{m}_t, \mathbf{S}_t)$	PSDA: $q(\mathbf{y}) = \mathcal{V}(\mathbf{y}|\mathbf{m}_t, r_t)$
$\boldsymbol{\Lambda}_t = \gamma_t \mathbf{W}^{-1} + \boldsymbol{\Lambda}_{t-1}$ $\boldsymbol{\eta}_t = \gamma_t \mathbf{W}^{-1}\mathbf{x}_t + \boldsymbol{\eta}_{t-1}$ $\mathbf{S}_t = \boldsymbol{\Lambda}_t^{-1}$, $\mathbf{S}_0 = \mathbf{B}$ $\mathbf{m}_t = \mathbf{S}_t\boldsymbol{\eta}_t$, $\mathbf{m}_0 = \boldsymbol{\mu}$	$\boldsymbol{\eta}_t = w\gamma_t\mathbf{x}_t + \boldsymbol{\eta}_{t-1}$ $r_t = \|\boldsymbol{\eta}_t\|$, $r_0 = b$ $\mathbf{m}_t = \boldsymbol{\eta}_t \,/\, r_t$, $\mathbf{m}_0 = \boldsymbol{\mu}$

Table 4.1: Update equations for the full-rank PLDA (4.1) and PSDA (4.2) at the time step t. The speaker index is omitted for clarity.

To detect new speakers we introduce an extra class corresponding to an unknown speaker. For this class, the posterior for the speaker identity variable is equal to the prior.

Algorithm 2 outlines the time step t of the proposed algorithms.

Algorithm 2 Proposed algorithm (time step t)

```
$\boldsymbol{\gamma}_t \equiv q(\mathbf{z}_t) \leftarrow$ Eq. (4.4)        ▷ Cluster membership probabilities
$q(\mathbf{y}_k) \leftarrow$ Table 4.1                                     ▷ Update clusters
$k \leftarrow \arg\max_i \gamma_{t,i}$                                     ▷ Find the most probable cluster
if $k = K + 1$ then                                                        ▷ New class is detected
    $K \leftarrow K + 1$                                                   ▷ Increment the total number of clusters
end if
```

The advantage of the proposed algorithm over Algorithm 1 is that it uses soft decisions for updating clusters. This makes the algorithm more robust to classification errors.

As a baseline for our experiments, we use Algorithm 1 with cosine similarity scoring.

4.4 Experiments

In this section, we analyze the performance of several back-end scoring models in the multi-enrollment scenario. First, we report results for a rarely investigated speaker verification scenario, *i.e.*, where the number of enrollment and test segments *varies* within an evaluation protocol. Next, we apply the proposed Algorithm 2 for the online speaker diarization task. To support reproducible research, we make the code and evaluation protocols publicly available.

We used open-source speaker embedding extractors in order to make our experiments reproducible. We decided to stick to the following systems: SpeechBrain [153], BUT model [152], and CLOVA [154]. Due to space limitations, we report results only for SpeechBrain, while other results can be found at the project repository[1].

4.4.1 Multi-enrollment verification

In this section, we compare different scoring back-ends in multi-enrollment speaker verification scenario. Specifically, we investigate calibration properties of the verification scores in the case where the number of enrollment and test segment varies within an evaluation protocol.

Experimental setup. We created several custom evaluation protocols from the VoxCeleb1 test set [33]. Specifically, we generated four trial lists with configurations $(1, 1)$, $(3, 1)$, $(10, 1)$, and $(3, 3)$, where the notation (#enrollments, #tests) represents the number of enrollment or test segments in a single trial. In addition, we combined all the trial lists to get the *pooled* protocol. The idea behind it is to reveal the robustness of scoring back-ends to the number of enrollment segments. To exclude the effect of utterance duration, the recordings were cropped to 2 seconds before extracting embeddings.

We compared several different scoring variants: cosine similarity with embedding averaging (CSEA) or score averaging (CSSA), PSDA [145], and three versions of PLDA with spherical, diagonal, and full covariance matrices. For PLDA and PSDA by-the-book scoring was used. The VoxCeleb1 dev set [33] was used for training the back-ends. We used two performance metrics: the equal error rate (EER) and the minimum normalized detection cost function (minDCF) with $P_{\text{Tar}} = 0.01$ [155].

[1] `https://github.com/sholokhovalexey/online-speaker-clustering`

Results. Figure 4.3 demonstrates the distribution of verification scores for different numbers of enrollment segments. One can see considerable distribution shifts for the target scores computed with CSEA. To be precise, a large variation of EER thresholds clearly makes each one sub-optimal for the other protocols. In contrast, the EER thresholds seem to be more stable for other scoring back-ends, even despite large differences in distribution means and variances for PLDA and PSDA. These observations are supported by objective metrics presented in Table 4.2. Despite low EERs for each protocol individually, the performance of CSEA degrades significantly on the pooled protocol. In contrast, CSSA does not suffer from this problem, however, it has higher error rates on the other protocols. Finally, PLDA and PSDA perform the best, overall, handling well all the cases. They also have very similar metrics and distributions of scores. These results are also in line with findings in [126] where spherical PLDA outperformed cosine similarity in the household speaker recognition task. Note that CSEA, CSSA, and sph-PLDA have exactly the same metrics in the $(1, 1)$ protocol because sph-PLDA is equivalent to cosine scoring. Another observation is that models with more parameters, diag- and full-PLDA, have comparable performance to sph-PLDA. This motivates choosing sph-PLDA as a simpler and faster alternative.

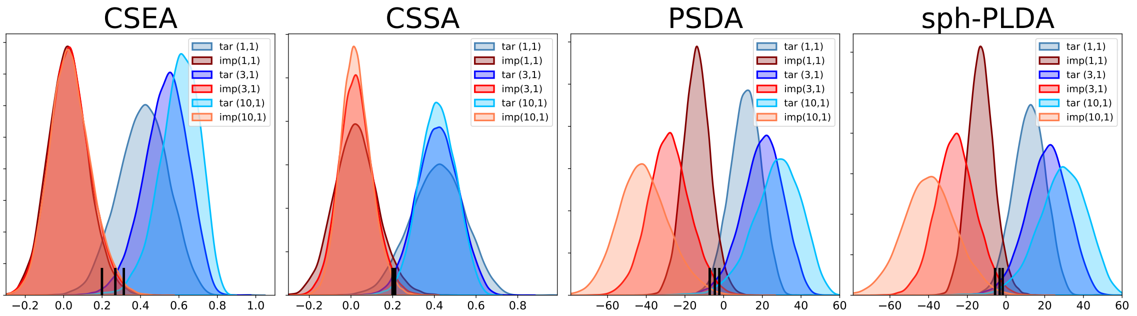


Figure 4.3: Distributions of target and impostor scores for different numbers of enrollment segments: 1, 3, and 10. Short black vertical lines represent EER thresholds.

Back-end	Evaluation protocol				
	(1, 1)	(3, 1)	(10, 1)	(3, 3)	pooled
CSEA	4.98	1.65	0.83	0.17	2.85 / 0.206
CSSA	4.98	1.79	1.02	0.37	2.05 / 0.228
PSDA	4.85	1.55	0.78	0.13	2.08 / 0.172
sph-PLDA	4.98	1.59	0.78	0.14	1.99 / 0.170
diag-PLDA	4.95	1.62	0.78	0.13	1.98 / 0.169
full-PLDA	4.74	1.79	1.08	0.20	2.06 / 0.201

Table 4.2: Comparison of the speaker verification performance for different scoring back-ends in terms of EER, %. The last column shows minDCF as well. SpeechBrain embeddings were used.

It should be noted that, unlike this study, PLDA model studied in [136] was not robust to the number of enrollment utterances. This probably can be explained by the nature of i-vector distribution which is different from the distribution of large-margin embeddings.

4.4.2 Online speaker diarization

In this section, we describe experiments on online speaker diarization. We used the same PLDA and PSDA models as for the previous experiments.

Experimental setup. We used two popular datasets of multi-speaker recordings: AMI [34], and VoxConverse [156]. Again, due to space limitations, we report only the results for the first one, while similar observations were made for the VoxConverse.

We used the development/evaluation split for the AMI corpus from [152][2]. The development set was used for tuning the hyper-parameters of the back-end models, pretrained on the VoxCeleb data.

For AMI, the evaluation was performed on Mix-Headset channel. We extracted embeddings from segments of length 2.0 sec with 1.0 sec overlap within the boundaries obtained by the ground-truth annotation. These embeddings were sequentially processed by several online clustering algorithms, producing the output annotation. We did not use any special heuristics for handling segments with overlapped speakers, thus one speaker was assigned to each segment.

We compared three versions of Algorithm 1: with CSEA, CSSA, and PLDA scoring. Also, we evaluated two versions of the proposed Algorithm 2, with sph-PLDA and PSDA models. All of the algorithms have at least one hyper-parameter (*e.g.* decision threshold) that was tuned on the development split.

Results. For the evaluation metrics, we use the diarization error rate (DER) [157] and Jaccard error rate (JER) [158]. The forgiveness collar was set to 0.25, and overlapped speech regions were excluded from evaluation for DER, however, JER is calculated with no forgiveness collar and includes overlapped speech [158]. Table 4.3 provides the evaluation results. Unlike the previous experiment on speaker verification, we found that sequential

Clustering back-end	DER, %	JER, %
Algorithm 1 w/ CSEA	3.63	25.20
Algorithm 1 w/ CSSA	3.67	26.33
Algorithm 1 w/ sph-PLDA	6.32	28.33
Algorithm 2 w/ PSDA	3.34	24.47
Algorithm 2 w/ sph-PLDA	3.32	25.21

Table 4.3: Online speaker diarization with SpeechBrain embeddings.

[2] https://github.com/BUTSpeechFIT/AMI-diarization-setup

PLDA scoring performs worse than cosine. As was discussed in [159] and [160], this probably can be explained by an inadequate assumption of statistical independence of the enrollment segments, which affects the score calibration. According to the theoretical model, enrollment segments are independent draws from the within-speaker distribution, while in diarization it is clearly not the case because of a shared acoustic environment and recording channel. However, unlike i-vector embeddings considered in [135, 159], this effect seems to be less evident for large-margin embeddings. Apparently, PLDA suffers from this effect only in diarization, while yielding adequate score calibration when embeddings are less statistically dependent, as in our previous experiment.

At the same time, the proposed clustering algorithm which uses the same PLDA model delivers lower error rates than Algorithm 1. In the future, we plan to further investigate the properties of this algorithm in other applications such as household speaker recognition, where speech utterances are also processed sequentially.

4.5 Conclusion

This chapter studies the properties of popular scoring back-ends suitable for large-margin speaker embeddings, with a particular focus on multi-enrollment speaker verification. Our experiments with the state-of-the-art embeddings revealed shortcomings of cosine scoring in the multi-enrollment scenario. To address this, we advocate for using the spherical PLDA that has several attractive properties: absence of numerical instabilities specific to PSDA due to Bessel functions; better performance, comparable computational complexity, and equivalence to cosine scoring in a special case. Also, we introduced a simple online clustering algorithm that uses the advantages of PLDA and PSDA for the multi-enrollment scenario. Empirical evaluation of the online speaker diarization showed superior performance of the proposed algorithm.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This study presents advancements in speaker anonymization within the framework of the Voice Privacy Challenge 2024 (VPC2024), emphasizing the use of disentanglement learning techniques to enhance privacy protection while preserving speech utility. The primary contributions focus on modifying baseline systems B3 and B5 from VPC2024 to improve their anonymization capabilities. For B3, emotion embeddings and advanced speaker embeddings, such as WavLM and ECAPA2, were incorporated alongside prosody manipulation, yielding a 15.5% improvement in Unweighted Average Recall (UAR) on emotion recognition for the IEMOCAP dataset. The B5 modifications, utilizing Mean Reversion and additive white Gaussian noise (AWGN) on prosody, achieved a 32.2% increase in Equal Error Rate (EER) on the Librispeech dataset, securing third place in privacy rankings at VPC2024. Additionally, the study adapts NaturalSpeech3 FACodec as a disentanglement-based approach for Speaker Anonymization, which demonstrated promising results compared to traditional β-VAE methods.

In the context of household speaker recognition, the study develops a probabilistic back-end that enhances online speaker recognition and clustering tasks. This model, tested on VoxCeleb1 and AMI datasets, achieved a 4% improvement in EER and an 8% reduction in Diarization Error Rate (DER), respectively. These results support the potential of integrating disentanglement-based methods in speaker anonymization frameworks within smart home environments, addressing both privacy and utility challenges. The research underscores the importance of disentangling linguistic and paralinguistic features for effective privacy preservation and lays the groundwork for future applications of disentanglement learning in smart home devices.

5.2 Room for Improvement

The goal of this thesis is to develop reliable speaker anonymization approaches by incorporating disentanglement learning techniques and to ensure their applicability to real-world scenarios. It is crucial to identify areas for improvement and address current weaknesses to enhance the effectiveness and robustness of these methods.

According to the results presented in Chapter 3, current disentanglement models do not achieve complete disentanglement, as there are information leaks between the supposedly independent components, such as speaker and emotion features. Additionally, despite significant improvements in emotion recognition performance on anonymized speech, there remains leakage of speaker information within the emotion embeddings, which compromises the anonymity of the speaker and leads to vulnerabilities in speaker anonymization.

In Chapter 4, we further explored a household speaker recognition scenario that involves handling sensitive user information in the form of speech data being transmitted to the cloud for processing. This process introduces several vulnerabilities to potential hacker attacks, as illustrated in Figure 1.1. To mitigate these risks and protect users' privacy, we propose incorporating the speaker anonymization approach based on disentanglement learning introduced in Chapter 3.

5.3 Future work

Future work will aim on areas of improvement highlighted in Section 5.2. Specifically, efforts will focus on achieving better disentanglement in models to prevent information leakage between independent components like speaker and emotion features. Enhancing the robustness of emotion embeddings to conceal speaker identity without compromising emotion recognition performance will be a key objective. Additionally, we plan to implement the speaker anonymization approach for household speaker recognition systems to better protect sensitive user information during cloud processing. This includes developing methods to securely handle speech data and mitigate potential vulnerabilities to hacker attacks. By addressing these areas, we hope to advance the field of speaker anonymization and contribute to more secure and privacy-preserving speech processing technologies.

Bibliography

[1] G. Tur and R. De Mori, "Spoken language understanding: Systems for extracting semantic information from speech," *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, Mar. 2011. DOI: 10.1002/9781119992691.

[2] A. McStay, *Emotional AI: The Rise of Empathic Media.* May 2018, ISBN: 9781473971103. DOI: 10.4135/9781526451293.

[3] Seyed *et al.*, "The 2016 nist speaker recognition evaluation," en, Interspeech 2017, Stockholm, -1, Aug. 2017. [Online]. Available: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=922849.

[4] G. Wei, Y. Zhang, H. Min, and Y. Xu, "End-to-end speaker identification research based on multi-scale sincnet and cgan," *Neural Comput. Appl.*, vol. 35, no. 30, pp. 22 209–22 222, Aug. 2023, ISSN: 0941-0643. DOI: 10.1007/s00521-023-08906-1. [Online]. Available: https://doi.org/10.1007/s00521-023-08906-1.

[5] J. Ziegeldorf, O. Morchon, and K. Wehrle, "Privacy in the internet of things: Threats and challenges," *Security and Communication Networks*, vol. 7, Dec. 2014. DOI: 10.1002/sec.795.

[6] R. Ko and R. Choo, *The Cloud Security Ecosystem: Technical, Legal, Business and Management Issues*, 1st. Syngress Publishing, 2015, ISBN: 0128015950.

[7] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," PMLR, 2017, pp. 1273–1282.

[8] P. Kairouz *et al.*, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[9] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," *Commun. ACM*, vol. 21, no. 2, pp. 120–126, Feb. 1978, ISSN: 0001-0782. DOI: 10.1145/359340.359342. [Online]. Available: https://doi.org/10.1145/359340.359342.

[10] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, Jan. 2019, ISSN: 2157-6904. DOI: 10.1145/3298981. [Online]. Available: https://doi.org/10.1145/3298981.

[11] N. Tomashenko *et al.*, "Introducing the voiceprivacy initiative," in *Interspeech 2020*, 2020, pp. 1693–1697. DOI: 10.21437/Interspeech.2020-1333.

[12] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[13] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28, Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf.

[14] H. Kim and A. Mnih, "Disentangling by factorising," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Jul. 2018, pp. 2649–2658. [Online]. Available: https://proceedings.mlr.press/v80/kim18b.html.

[15] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*, PMLR, 2019, pp. 5210–5219.

[16] J.-c. Chou, C.-c. Yeh, H.-y. Lee, and L.-S. Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," Sep. 2018, pp. 501–505. DOI: 10.21437/Interspeech.2018-1830.

[17] N. A. Tomashenko *et al.*, "The voiceprivacy 2020 challenge evaluation plan," *ArXiv*, vol. abs/2205.07123, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211550970.

[18] M. Todisco *et al.*, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," Sep. 2019, pp. 1008–1012. DOI: 10.21437/Interspeech.2019-2249.

[19] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, "Clova baseline system for the VoxCeleb speaker recognition challenge 2020," *arXiv:2009.14153*, 2020.

[20] S. Pascual, J. Serrà, and A. Bonafonte, "Towards generalized speech enhancement with generative adversarial networks," in *Interspeech 2019*, 2019, pp. 1791–1795. DOI: 10.21437/Interspeech.2019-2688.

[21] A. Polyak *et al.*, "Speech resynthesis from discrete disentangled self-supervised representations," Aug. 2021, pp. 3615–3619. DOI: 10.21437/Interspeech.2021-475.

[22] F. Fang *et al.*, "Speaker anonymization using x-vector and neural waveform models," Sep. 2019, pp. 155–160. DOI: 10.21437/SSW.2019-28.

[23] N. Kuzmin, H.-T. Luong, J. Yao, L. Xie, K. A. Lee, and E.-S. Chng, "Ntu-npu system for voice privacy 2024 challenge," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 72–79. DOI: 10.21437/SPSC.2024-13.

[24] N. Tomashenko *et al.*, "The VoicePrivacy 2024 challenge evaluation plan," 2024. arXiv: 2404.02677 [eess.AS].

[25] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022, ISSN: 1941-0484. DOI: 10.1109/jstsp.2022.3188113. [Online]. Available: http://dx.doi.org/10.1109/JSTSP.2022.3188113.

[26] J. Thienpondt and K. Demuynck, "Ecapa2: A hybrid neural network architecture and training strategy for robust speaker embeddings," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023.

[27] C. Busso *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, Dec. 2008. DOI: 10.1007/s10579-008-9076-6.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[29] Z. Ju *et al.*, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," *ArXiv*, vol. abs/2403.03100, 2024.

[30] H. Lu, D. Wang, X. Wu, Z. Wu, X. Liu, and H. Meng, "Disentangled speech representation learning for one-shot cross-lingual voice conversion using ß-vae," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 814–821.

[31] A. Sholokhov, N. Kuzmin, K. A. Lee, and E. S. Chng, "Probabilistic back-ends for online speaker recognition and clustering," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[32] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *The Speaker and Language Recognition Workshop (Odyssey 2010)*, 2010, paper 34.

[33] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *INTERSPEECH*, 2017, pp. 2616–2620.

[34] I. McCowan *et al.*, "The AMI meeting corpus," *Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, pp. 137–140, 2005.

[35] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the mcadams coefficient," in *Proc. Interspeech 2021*, 2021, pp. 1099–1103. DOI: `10.21437/Interspeech.2021-1070`.

[36] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018, pp. 5329–5333.

[37] F. Fang *et al.*, "Speaker anonymization using x-vector and neural waveform models," in *Speech Synthesis Workshop*, 2019. [Online]. Available: `https://api.semanticscholar.org/CorpusID:173188773`.

[38] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 4725–4729.

[39] P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," Ph.D. dissertation, Université de Lorraine, 2023.

[40] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and T. Vu, "Prosody is not identity: A speaker anonymization approach using prosody cloning," Jun. 2023, pp. 1–5. DOI: `10.1109/ICASSP49357.2023.10096607`.

[41] S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, "Anonymizing speech with generative adversarial networks to preserve speaker privacy," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2023, pp. 912–919.

[42] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The gdpr & speech data: Reflections of legal and technology communities, first steps towards a common understanding," in *Interspeech 2019*, ser. interspeech_2019, ISCA, Sep. 2019. DOI: `10.21437/interspeech.2019-2647`. [Online]. Available: `http://dx.doi.org/10.21437/Interspeech.2019-2647`.

[43] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[44] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. [Online]. Available: `https://api.semanticscholar.org/CorpusID:2191379`.

[45] M. Ravanelli *et al.*, *Speechbrain: A general-purpose speech toolkit*, 2021. arXiv: `2106.04624 [eess.AS]`. [Online]. Available: `https://arxiv.org/abs/2106.04624`.

[46] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *INTERSPEECH*, 2020, pp. 3830–3834.

[47] J. Qian *et al.*, "Voicemask: Anonymize and sanitize voice input on mobile devices," *CoRR*, vol. abs/1711.11460, 2017. arXiv: `1711.11460`. [Online]. Available: `http://arxiv.org/abs/1711.11460`.

[48] T. Vaidya and M. Sherr, "You talk too much: Limiting privacy exposure via voice input," in *2019 IEEE Security and Privacy Workshops (SPW)*, 2019, pp. 84–91. DOI: `10.1109/SPW.2019.00026`.

[49] H. Kai, S. Takamichi, S. Shiota, and H. Kiya, "Lightweight voice anonymization based on data-driven optimization of cascaded voice modification modules," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 560–566. DOI: `10.1109/SLT48900.2021.9383535`.

[50] C. O. Mawalim, S. Okada, and M. Unoki, "Speaker anonymization by pitch shifting based on time-scale modification," in *2nd Symposium on Security and Privacy in Speech Communication*, 2022, pp. 35–42. DOI: `10.21437/SPSC.2022-7`.

[51] I.-C. Yoo, K. Lee, S. Leem, H. Oh, B. Ko, and D. Yook, "Speaker anonymization for personal information protection using voice conversion techniques," *IEEE Access*, vol. 8, pp. 198 637–198 645, 2020. DOI: `10.1109/ACCESS.2020.3035416`.

[52] S. Ahmed, A. R. Chowdhury, K. Fawaz, and P. Ramanathan, "Preech: A system for Privacy-Preserving speech transcription," in *29th USENIX Security Symposium (USENIX Security 20)*, USENIX Association, Aug. 2020, pp. 2703–2720, ISBN: 978-1-939133-17-5. [Online]. Available: `https://www.usenix.org/conference/usenixsecurity20/presentation/ahmed-shimaa`.

[53] R. Aloufi, H. Haddadi, and D. Boyle, "Privacy-preserving voice analysis via disentangled representations," in *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, ser. CCSW'20, New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–14, ISBN: 9781450380843. DOI: `10.1145/3411495.3421355`. [Online]. Available: `https://doi.org/10.1145/3411495.3421355`.

[54] D. K. Singh, G. P. Prajapati, and H. A. Patil, "Voice privacy using time-scale and pitch modification," *SN Comput. Sci.*, vol. 5, no. 2, Jan. 2024. DOI: `10.1007/s42979-023-02549-8`. [Online]. Available: `https://doi.org/10.1007/s42979-023-02549-8`.

[55] B. M. L. Srivastava *et al.*, "Design choices for x-vector based speaker anonymization," in *Interspeech 2020*, 2020, pp. 1713–1717. DOI: `10.21437/Interspeech.2020-2692`.

[56] H. Turner, G. Lovisotto, and I. Martinovic, "Speaker anonymization with distribution-preserving x-vector generation for the voiceprivacy challenge 2020," *arXiv preprint arXiv:2010.13457*, 2020.

[57] C. O. Mawalim, K. Galajit, J. Karnjana, and M. Unoki, "X-vector singular value modification and statistical-based decomposition with ensemble regression modeling for speaker anonymization system," in *Interspeech 2020*, 2020, pp. 1703–1707. DOI: `10.21437/Interspeech.2020-1887`.

[58] F. M. Espinoza-Cuadros, J. M. Perero-Codosero, J. Antón-Martín, and L. A. Hernández-Gómez, "Speaker de-identification system using autoencoders and adversarial training," *arXiv preprint arXiv:2011.04696*, 2020.

[59] A. S. Shamsabadi *et al.*, "Differentially private speaker anonymization," *Proceedings on Privacy Enhancing Technologies*, 2023.

[60] H. Turner, G. Lovisotto, S. Eberz, and I. Martinovic, *I'm hearing (different) voices: Anonymous voices to protect user privacy*, 2022. arXiv: `2202.06278 [cs.CR]`. [Online]. Available: `https://arxiv.org/abs/2202.06278`.

[61] J. Deng, F. Teng, Y. Chen, X. Chen, Z. Wang, and W. Xu, "V-Cloak: Intelligibility-, naturalness- & Timbre-Preserving Real-Time voice anonymization," in *32nd USENIX Security Symposium (USENIX Security 23)*, Anaheim, CA: USENIX Association, Aug. 2023, pp. 5181–5198, ISBN: 978-1-939133-37-3. [Online]. Available: `https://www.usenix.org/conference/usenixsecurity23/presentation/deng-jiangyi-v-cloak`.

[62] P. O'Reilly, A. Bugler, K. Bhandari, M. Morrison, and B. Pardo, "Voiceblock: Privacy through real-time adversarial attacks with audio-to-audio models," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., 2022, pp. 30 058–30 070. [Online]. Available: `https://proceedings.neurips.cc/paper_files/paper/2022/file/c204d12afa0175285e5aac65188808b4-Paper-Conference.pdf`.

[63] N. Tomashenko *et al.*, *The voiceprivacy 2022 challenge evaluation plan*, 2022. arXiv: `2203.12468 [eess.AS]`. [Online]. Available: `https://arxiv.org/abs/2203.12468`.

[64] M. Tran and M. Soleymani, "A speech representation anonymization framework via selective noise perturbation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: `10.1109/ICASSP49357.2023.10095173`.

[65] X. Wang and J. Yamagishi, "Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis," Sep. 2019, pp. 1–6. DOI: `10.21437/SSW.2019-1`.

[66] P.-G. Noé, J.-F. Bonastre, D. Matrouf, N. Tomashenko, A. Nautsch, and N. Evans, "Speech pseudonymisation assessment using voice similarity matrices," in *Interspeech 2020*, 2020, pp. 1718–1722. DOI: `10.21437/Interspeech.2020-2720`.

[67] P.-G. Noé *et al.*, "Towards a unified assessment framework of speech pseudonymisation," *Computer Speech & Language*, vol. 72, p. 101 299, 2022, ISSN: 0885-2308. DOI: `https://doi.org/10.1016/j.csl.2021.101299`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0885230821001005`.

[68] A. Nautsch, "Survey talk: Preserving privacy in speaker and speech characterisation," in *Interspeech 2019*, 2019.

[69] J. Qian, F. Han, J. Hou, C. Zhang, Y. Wang, and X.-Y. Li, "Towards privacy-preserving speech data publishing," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, Honolulu, HI, USA: IEEE Press, 2018, pp. 1079–1087. DOI: `10.1109/INFOCOM.2018.8486250`. [Online]. Available: `https://doi.org/10.1109/INFOCOM.2018.8486250`.

[70] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 2802–2806.

[71] M. Maouche, B. M. L. Srivastava, N. Vauquier, A. Bellet, M. Tommasi, and E. Vincent, “Enhancing speech privacy with slicing,” in *Interspeech 2022 - Human and Humanizing Speech Technology*, Incheon, South Korea, Sep. 2022. [Online]. Available: `https://inria.hal.science/hal-03369137`.

[72] B. M. L. Srivastava *et al.*, “Privacy and utility of x-vector based speaker anonymization,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 30, pp. 2383–2395, Jul. 2022, ISSN: 2329-9290. DOI: `10.1109/TASLP.2022.3190741`. [Online]. Available: `https://doi.org/10.1109/TASLP.2022.3190741`.

[73] J. Kahn *et al.*, “Libri-light: A benchmark for asr with limited or no supervision,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2020, pp. 7669–7673. DOI: `10.1109/icassp40776.2020.9052942`. [Online]. Available: `http://dx.doi.org/10.1109/ICASSP40776.2020.9052942`.

[74] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds., Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2236–2246. DOI: `10.18653/v1/P18-1208`. [Online]. Available: `https://aclanthology.org/P18-1208`.

[75] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english,” *PLOS ONE*, vol. 13, no. 5, pp. 1–35, May 2018. DOI: `10.1371/journal.pone.0196391`. [Online]. Available: `https://doi.org/10.1371/journal.pone.0196391`.

[76] S. Mariooryad, R. Lotfian, and C. Busso, “Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora,” in *Interspeech*, 2014. [Online]. Available: `https://api.semanticscholar.org/CorpusID:973607`.

[77] G. Sharma, A. Dhall, and J. Cai, “ Audio-Visual Automatic Group Affect Analysis,” *IEEE Transactions on Affective Computing*, vol. 14, no. 02, pp. 1056–1069, Apr. 2023, ISSN: 1949-3045. DOI: `10.1109/TAFFC.2021.3104170`. [Online]. Available: `https://doi.ieeecomputersociety.org/10.1109/TAFFC.2021.3104170`.

[78] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 920–924.

[79] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, pp. 377–390, 2014. [Online]. Available: `https://api.semanticscholar.org/CorpusID:14369452`.

[80] S. Haq and P. J. B. Jackson, "Speaker-dependent audio-visual emotion recognition," in *Auditory-Visual Speech Processing*, 2009, pp. 53–58.

[81] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Interspeech*, 2005. [Online]. Available: `https://api.semanticscholar.org/CorpusID:13920681`.

[82] H. Zen *et al.*, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech 2019*, 2019, pp. 1526–1530. DOI: `10.21437/Interspeech.2019-2441`.

[83] K. Ito and L. Johnson, *The lj speech dataset*, `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[84] J. Yamagishi, C. Veaux, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019. [Online]. Available: `https://api.semanticscholar.org/CorpusID:213060286`.

[85] D. Snyder, G. Chen, and D. Povey, *Musan: A music, speech, and noise corpus*, 2015. arXiv: `1510.08484 [cs.SD]`. [Online]. Available: `https://arxiv.org/abs/1510.08484`.

[86] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224, 2017. [Online]. Available: `https://api.semanticscholar.org/CorpusID:23138179`.

[87] D. Povey *et al.*, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Interspeech 2018*, 2018, pp. 3743–3747. DOI: `10.21437/Interspeech.2018-1417`.

[88] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Interspeech 2015*, 2015, pp. 3214–3218. DOI: `10.21437/Interspeech.2015-647`.

[89] D. Povey *et al.*, “The kaldi speech recognition toolkit,” *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Jan. 2011.

[90] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[91] S. McAdams, “Spectral fusion, spectral parsing and the formation of auditory images,” M.S. thesis, Stanford University, Stanford, California, May 1984. [Online]. Available: `https://ccrma.stanford.edu/files/papers/stanm22.pdf`.

[92] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and T. Vu, “Prosody is not identity: A speaker anonymization approach using prosody cloning,” Jun. 2023, pp. 1–5. DOI: `10.1109/ICASSP49357.2023.10096607`.

[93] Y. Wang *et al.*, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *International conference on machine learning*, PMLR, 2018, pp. 5180–5189.

[94] S. Watanabe, T. Hori, S. Kim, J. Hershey, and T. Hayashi, “Hybrid ctc/attention architecture for end-to-end speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, pp. 1–1, Oct. 2017. DOI: `10.1109/JSTSP.2017.2763455`.

[95] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, “Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding,” in *International Conference on Machine Learning*, PMLR, 2022, pp. 17 627–17 643.

[96] H. Liu, X. Gu, and D. Samaras, “Wasserstein gan with quadratic transport cost,” *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4831–4840, 2019.

[97] Y. Ren *et al.*, “Fastspeech 2: Fast and high-quality end-to-end text to speech,” in *International Conference on Learning Representations*, 2021. [Online]. Available: `https://openreview.net/forum?id=piLPYqxtWuA`.

[98] F. Lux, J. Koch, A. Schweitzer, and N. Thang Vu, “The ims toucan system for the blizzard challenge 2021,” in *The Blizzard Challenge 2021*, 2021, pp. 14–19. DOI: `10.21437/Blizzard.2021-2`.

[99] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A.-r. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021. [Online]. Available: `https://api.semanticscholar.org/CorpusID:235421619`.

[100] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023, Featured Certification, Reproducibility Certification, ISSN: 2835-8856. [Online]. Available: `https://openreview.net/forum?id=ivCd8z8zR2`.

[101] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2100–2104. DOI: `10.23919/EUSIPCO.2018.8553236`.

[102] H. Lu, D. Wang, X. Wu, Z. Wu, X. Liu, and H. Meng, "Disentangled speech representation learning for one-shot cross-lingual voice conversion using β-vae," Jan. 2023, pp. 814–821. DOI: `10.1109/SLT54892.2023.10022787`.

[103] Y. Bengio, "Deep learning of representations: Looking forward," in *International conference on statistical language and speech processing*, Springer, 2013, pp. 1–37.

[104] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2414–2423. DOI: `10.1109/CVPR.2016.265`.

[105] F. Locatello, S. Bauer, M. Lucic, S. Gelly, B. Scholkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *International Conference on Machine Learning*, 2018. [Online]. Available: `https://api.semanticscholar.org/CorpusID:54089884`.

[106] F. Locatello *et al.*, "A sober look at the unsupervised learning of disentangled representations and their evaluation," *J. Mach. Learn. Res.*, vol. 21, no. 1, Jan. 2020, ISSN: 1532-4435.

[107] F. Locatello, B. Poole, G. Rätsch, B. Scholkopf, O. Bachem, and M. Tschannen, "Weakly-supervised disentanglement without compromises," in *International Conference on Machine Learning*, 2020. [Online]. Available: `https://api.semanticscholar.org/CorpusID:211066424`.

[108] Y.-J. Luo, K. Agres, and D. Herremans, “Learning disentangled representations of timbre and pitch for musical instrument sounds using gaussian mixture variational autoencoders,” *Proceedings of International Society for Music Information Retrieval Conference*, pp. 746–753, 2019.

[109] A. van den Oord, O. Vinyals, and k. kavukcuoglu koray, “Neural discrete representation learning,” in *Advances in Neural Information Processing Systems*, I. Guyon *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: `https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf`.

[110] M. Baas and H. Kamper, “Disentanglement in a gan for unconditional speech synthesis,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1324–1335, 2023. [Online]. Available: `https://api.semanticscholar.org/CorpusID:259341784`.

[111] J. Williams, “Learning disentangled speech representations,” Ph.D. dissertation, University of Edinburgh, 2022.

[112] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, “Speechtokenizer: Unified speech tokenizer for speech language models,” in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: `https://openreview.net/forum?id=AF9Q8Vip84`.

[113] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML’15, Lille, France: JMLR.org, 2015, pp. 1180–1189.

[114] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” *CoRR*, vol. abs/1711.00937, 2017. arXiv: `1711.00937`. [Online]. Available: `http://arxiv.org/abs/1711.00937`.

[115] K. Qian, Y. Zhang, S. Chang, D. Cox, and M. Hasegawa-Johnson, “Unsupervised speech decomposition via triple information bottleneck,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML’20, JMLR.org, 2020.

[116] F. Lux *et al.*, “The IMS Toucan System for the Blizzard Challenge 2023,” in *Blizzard Challenge Workshop*, ISCA Speech Synthesis SIG, 2023.

[117] S. A. Zahorian and H. Hu, “A spectral/temporal method for robust fundamental frequency tracking,” *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.

[118] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," 2021. arXiv: `2110.13900 [cs.CL]`.

[119] J. Thienpondt and K. Demuynck, "Speaker embeddings with weakly supervised voice activity detection for efficient speaker diarization," May 2024.

[120] J. Wagner *et al.*, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis &amp; Machine Intelligence*, vol. 45, no. 09, pp. 10 745–10 759, Sep. 2023, ISSN: 1939-3539. DOI: `10.1109/TPAMI.2023.3263585`.

[121] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, pp. 471–483, 2019.

[122] S. Meyer, P. Tilli, F. Lux, P. Denisov, J. Koch, and N. T. Vu, "Cascade of phonetic speech recognition, speaker embeddings gan and multispeaker speech synthesis for the VoicePrivacy 2022 Challenge," in *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 2022.

[123] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," Oct. 2020, pp. 5036–5040. DOI: `10.21437/Interspeech.2020-3015`.

[124] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[125] J. Kahn *et al.*, "Libri-Light: A Benchmark for ASR with Limited or No Supervision," *arXiv e-prints*, arXiv:1912.07875, arXiv:1912.07875, Dec. 2019. DOI: `10.48550/arXiv.1912.07875`. arXiv: `1912.07875 [cs.CL]`.

[126] A. Sholokhov, X. Liu, M. Sahidullah, and T. Kinnunen, "Baselines and protocols for household speaker recognition," in *Odyssey*, 2022, pp. 185–192.

[127] Z. Tan, Y. Yang, E. Han, and A. Stolcke, "Improving speaker identification for shared devices by adapting embeddings to speaker subsets," in *ASRU*, 2021, pp. 1124–1131.

[128] J. Patino *et al.*, "Low-latency speaker spotting with online diarization and detection," in *Odyssey*, 2018, pp. 140–146.

[129] D. Liu and F. Kubala, "Online speaker clustering," in *ICASSP*, 2004, pp. 333–336.

[130] R. Aloni-Lavi, I. Opher, and I. Lapidot, "Incremental on-line clustering of speakers' short segments," in *Odyssey*, 2018, pp. 120–127.

[131] G. Wisniewski, H. Bredin, G. Gelly, and C. Barras, "Combining speaker turn embedding and incremental structure prediction for low-latency speaker diarization," in *INTERSPEECH*, 2017, pp. 3582–3586.

[132] T. Koshinaka, K. Nagatomo, and K. Shinoda, "Online speaker clustering using incremental learning of an ergodic hidden Markov model," *IEICE Trans. Inf. Syst.*, vol. 95-D, no. 10, pp. 2469–2478, 2012.

[133] W. Zhu and J. W. Pelecanos, "Online speaker diarization using adapted *i*-vector transforms," in *ICASSP*, 2016, pp. 5045–5049.

[134] G. Soldi, C. Beaugeant, and N. W. D. Evans, "Adaptive and online speaker diarization for meeting data," in *EUSIPCO*, 2015, pp. 2112–2116.

[135] P. Rajan, A. Afanasyev, V. Hautamaki, and T. Kinnunen, "From single to multiple enrollment *i*-vectors: Practical PLDA scoring variants for speaker verification," *Digital Signal Processing*, vol. 31, pp. 93–101, 2014, ISSN: 1051-2004.

[136] K. Lee, A. Larcher, C. H. You, B. Ma, and H. Li, "Multi-session PLDA scoring of *i*-vector for partially open-set speaker detection," in *INTERSPEECH*, 2013, pp. 3651–3655.

[137] M. H. Soni and A. Panda, "LDA-based speaker verification in multi-enrollment scenario using expected vector approach," in *ISCSLP*, 2021, pp. 1–5.

[138] C. Zeng, X. Wang, E. Cooper, X. Miao, and J. Yamagishi, "Attention back-end for automatic speaker verification with multiple enrollment utterances," in *ICASSP*, 2022, pp. 6717–6721.

[139] Q. Wang, K. A. Lee, and T. Liu, "Scoring of large-margin embeddings for speaker verification: Cosine or PLDA?" *arXiv:2204.03965*, 2022.

[140] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Odyssey*, Brno, Czech Republic, 2010, pp. 194–201.

[141] J. A. V. López, M. Díez, A. Varona, and E. Lleida, "Handling recordings acquired simultaneously over multiple channels with PLDA," in *INTERSPEECH*, 2013, pp. 2509–2513.

[142] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech & Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[143] Z. Peng, X. He, K. Ding, T. Lee, and G. Wan, "Unifying cosine and PLDA backends for speaker verification," *arXiv:2204.10523*, 2022.

[144] N. Kuzmin, I. Fedorov, and A. Sholokhov, “Magnitude-aware probabilistic speaker embeddings,” in *Odyssey*, 2022, pp. 1–8.

[145] N. Brümmer *et al.*, “Probabilistic spherical discriminant analysis: An alternative to PLDA for length-normalized embeddings,” in *INTERSPEECH*, 2022, pp. 1446–1450.

[146] K. V. Mardia and P. E. Jupp, *Directional statistics.* Wiley Online Library, 2000, vol. 2.

[147] P. A. Mansfield, Q. Wang, C. Downey, L. Wan, and I. Lopez-Moreno, “Links: A high-dimensional online clustering method,” *arXiv:1801.10123*, 2018.

[148] F. Valente and C. Wellekens, “Variational Bayesian speaker clustering,” in *Odyssey*, 2004, pp. 207–214.

[149] M. Díez, L. Burget, F. Landini, and J. Cernocký, “Analysis of speaker diarization based on Bayesian HMM with eigenvoice priors,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 355–368, 2020.

[150] A. Corduneanu and C. Bishop, “Variational Bayesian model selection for mixture distributions,” in *AISTATS*, 2001, pp. 27–34. [Online]. Available: `https://www.microsoft.com/en-us/research/publication/variational-bayesian-model-selection-for-mixture-distributions/`.

[151] C. M. Bishop, *Pattern Recognition and Machine Learning.* New York: Springer Science+Business Media, LLC, 2006.

[152] F. Landini, J. Profant, M. Díez, and L. Burget, “Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks,” *arXiv:2012.14952*, 2020.

[153] M. Ravanelli *et al.*, “SpeechBrain: A general-purpose speech toolkit,” *arXiv:2106.04624*, 2021.

[154] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, “Clova baseline system for the VoxCeleb speaker recognition challenge 2020,” *arXiv:2009.14153*, 2020.

[155] M. Przybocki and A. Martin, “NIST speaker recognition evaluation chronicles,” in *Odyssey*, 2004, pp. 15–22.

[156] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, “Spot the conversation: Speaker diarisation in the wild,” in *INTERSPEECH*, 2020, pp. 299–303.

[157] O. Galibert, "Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech," in *INTERSPEECH*, 2013, pp. 1131–1134.

[158] N. Ryant *et al.*, "The second DIHARD diarization challenge: Dataset, task, and baselines," in *INTERSPEECH*, 2019, pp. 978–982.

[159] A. McCree, G. Sell, and D. Garcia-Romero, "Extended variability modeling and unsupervised adaptation for PLDA speaker recognition," in *INTERSPEECH*, 2017, pp. 1552–1556.

[160] T. Stafylakis, P. Kenny, V. Gupta, and P. Dumouchel, "Compensation for interframe correlations in speaker diarization and recognition," in *ICASSP*, 2013, pp. 7731–7735.