AudioPrivacy: Parallel Audio Dataset for Speaker Profiling with Diverse Audio Types and Rich Attributes

Jiabei He, Yanzhe Zhang, Jiaming Zhou, Hui Wang, Haoqin Sun, Yong Qin*
College of Computer Science, Nankai University, Tianjin, China
hejiabei@mail.nankai.edu.cn

Abstract

Speech signals convey abundant speaker-related metadata, yet current privacy research predominantly focuses on identity-centric voiceprint protection, leaving sensitive Speaker Attribute Privacy (SAP) largely underexplored. This paper introduces AudioPrivacy [1], a large-scale Chinese dataset designed to systematically evaluate SAP leakage in realistic, everyday scenarios. Comprising 227.3 hours of audio from 1,000 speakers, it uniquely encompasses four parallel modalities: speech, singing, paralinguistic expressions, and non-vocal acoustic signals (e.g., footsteps). Annotated with 11 diverse attributes, including fine-grained physiological traits often overlooked in traditional corpora, AudioPrivacy enables a granular analysis of acoustic privacy risks. Our evaluations reveal significant leakage across multiple attributes, even when inferred from non-vocal signals. Furthermore, we demonstrate that state-of-the-art Multimodal Large Language Models (MM LLMs) can precisely profile speakers and exacerbate these risks, underscores the urgent need to rethink privacy-preserving mechanisms in the era of powerful audio foundation models.

1 Introduction

Speech signals serve as a multifaceted medium conveying not only linguistic content but also abundant speaker-related personal metadata (Cheng and Roedig, 2022). With the rapid evolution of AI-driven analysis, the extraction of such information has become alarmingly precise, intensifying global concerns regarding privacy leakage from acoustic signals (Bäckström, 2025). While contemporary research has predominantly prioritized voiceprint protection (Hanisch et al., 2025), this identity-centric paradigm is increasingly recognized as overly restrictive and insufficient to address the full spectrum of acoustic privacy risks (Miao et al., 2025). Consequently, there is an urgent mandate to pivot toward Speaker Attribute Privacy (SAP), a more nuanced framework that characterizes speakers through diverse sensitive traits. Compared to traditional identity threats, SAP leakage poses more immediate and tangible risks; as SAPs are directly tethered to real-world rights and social vulnerabilities, their unauthorized inference can act as a catalyst for systemic discrimination, unfair treatment, and social exclusion (Ren et al., 2023).

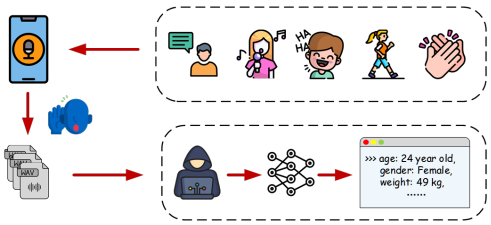


Figure 1: The scenario of speaker attribute privacy leakage from audio: the smartphone is corrupted and eavesdropped by the hacker to inferences the users' privacy.

The landscape of SAP extraction has been fundamentally reshaped by technological advancements. Early research treated speaker profiling (SP) as a series of isolated classification and regression tasks relying on hand-crafted features (Nautsch et al., 2019; Kalluri et al., 2020). However, the emergence of large-scale foundation models—such as wav2vec (Schneider et al., 2019; Baevski et al., 2020), Whisper (Radford et al., 2023), and WavLM (Chen et al., 2022), alongside the recent integration of multimodal large language models (MM LLMs) (Chu et al., 2024), has significantly amplified audio understanding capabilities. Inspired by the potential of CoLMbo (Baali et al., 2025), we argue that MM LLMs are redefining the SP task by unifying diverse audio modalities. Yet, despite these powerful analytical tools, research into the

*Corresponding Author.
[1] https://huggingface.co/datasets/krkr114/AudioPrivacy

Findings of the Association for Computational Linguistics: ACL 2026, pages 5735–5748
July 2-7, 2026

Corpus	Language	#Speakers	Supporting Attributes
TIMIT (Garofolo et al., 1993)	English	630	Gender, Age, American Dialect Regions
VCTK-RVA (Sheng et al., 2025)	English	110	Gender, Voice Attributes Label
LibriSpeech (Panayotov et al., 2015)	English	2484	Gender, Age, Reader Identity
Mixer 6 (Brandschain et al., 2010)	English	594	Gender, Age, Native Speaker Identity
SonicSet (of SonicSet, 2024)	English	2484	Gender, Movement Trajectory
WildElder (of WildElder, 2025)	Chinese	200	Gender, Age Group, Accent Intensity, Health Status
CN-Celeb (Fan et al., 2020)	Chinese	3000	Gender, Age, Occupation
KeSpeech (Tang et al., 2021)	Chinese	27237	Gender, Age, Region, Dialect Type
3D-Speaker (Zheng et al., 2023)	Chinese	10000+	Gender, Age, Dialect, Device Type
VoxCeleb1 (Nagrani et al., 2019)	Multilingual	1251	Gender, Name, Nationality, Regions
VoxCeleb2 (Chung et al., 2018)	Multilingual (145)	6112	Gender, Age, Ethnicity, Accent
NISP (Kalluri et al., 2021)	Multilingual (6)	345	Gender, Age, Height, Shoulder Width, Weight, Accent
HeightCeleb (Kacprzak and Kowalczyk, 2024)	Multilingual	1251	Gender, Age, Height
VoxBlink (Lin et al., 2024)	Multilingual	23137	Gender, Region, Language
VoxPopuli (Zhang et al., 2021)	Multilingual (23)	4295	Gender, Nationality, Native Language
SPGISpeech 2.0 (Grossman et al., 2025)	Multilingual (2)	5000+	Gender, Professional Identity, Emotional Tendency
M3SD (Wu et al., 2025)	Multilingual (10)	1000+	Gender, Dialogue Role
NIST SRE2021 (Sadjadi et al., 2022)	Multilingual	5000+	Gender, Nationality, Language Proficiency
AudioRole (Li et al., 2025)	Multilingual (2)	500+	Gender, Character Identity, Emotional State, Line Style
AudioPrivacy (Ours)	Chinese	1000	Gender, Age, Height, Weight, Shoulder Width, **Waist Circumference**, **Wrist Circumference**, Region, Education, **Shoe Size**, **Shoe Type**

Table 1: The key information of the dataset used for the Speaker Profiling (SP) task, along with the included speaker attributes. The unique attributes of AudioPrivacy are highlighted in **bold**.

systemic privacy risks they induce remains substantially limited, hindered by a critical mismatch between existing datasets and real-world conditions. As shown in Tab.1, most current corpora are collected in controlled studio environments and focus on restricted attribute sets, failing to reflect realistic eavesdropping scenarios where speakers' SAPs are exposed via ubiquitous personal devices.

To bridge this gap and provide a rigorous benchmark for the MM LLM era, we present AudioPrivacy, a large-scale Chinese dataset specifically designed for evaluating speaker privacy leakage risks in realistic settings. Comprising 227.3 hours of recordings from 1,000 speakers across 27 provincial-level regions, AudioPrivacy simulates authentic eavesdropping scenarios using personal smartphones. Distinct from previous datasets, it captures a holistic acoustic profile of the speaker by including four parallel audio types: speech (SPCH), singing (SING), paralinguistic signals (PRLG), and non-vocal acoustic signals (NVAS). Annotated with 11 diverse speaker attributes, AudioPrivacy is, to our knowledge, the most comprehensive dataset in the SP domain in terms of both audio diversity and attribute granularity. This multi-modal approach allows for the investigation of privacy leakage not only from voice but also from intrinsic biological proxies and extrinsic physical contexts.

The contributions of this work are as follows: (1) The AudioPrivacy Dataset, providing a largest-scale, realistic benchmark for fine-grained SAP research; (2) MM LLM-based Risk Assessment, demonstrating that advanced models can precisely profile speakers to achieve significantly heightened prediction accuracy in over half of the annotated SAPs; and (3) The RALG Metric, a novel measure for standardized privacy risk comparison across heterogeneous audio types and attributes. Our evaluations reveal a staggering reality: over half of the annotated SAPs pose a high risk of leakage under foundation models, highlighting the urgent need to rethink privacy protection in the AI era.

2 Related Work

Most datasets used for SP were originally designed for other tasks. As a result, in addition to common attributes such as gender and age, which are typically annotated in many speech tasks, the availability of additional attributes for SP varies, as does the focus of the audio collection. The commonly used datasets for SP can be classified based on their originally intended tasks. Representative datasets from each category are introduced, which can be broadly grouped into three types: Automatic Speech Recognition (ASR), speaker recognition, and SP.

2.1 Datasets for Automatic Speech Recognition

The TIMIT dataset (Garofolo et al., 1993), designed for automatic speech recognition (ASR), includes extensive speaker information, such as American dialect labels, height, race, education, and occupation. However, its audio is recorded in a controlled studio environment, limiting the diversity of recording conditions and equipment.

In contrast, KeSpeech (Tang et al., 2021) col-

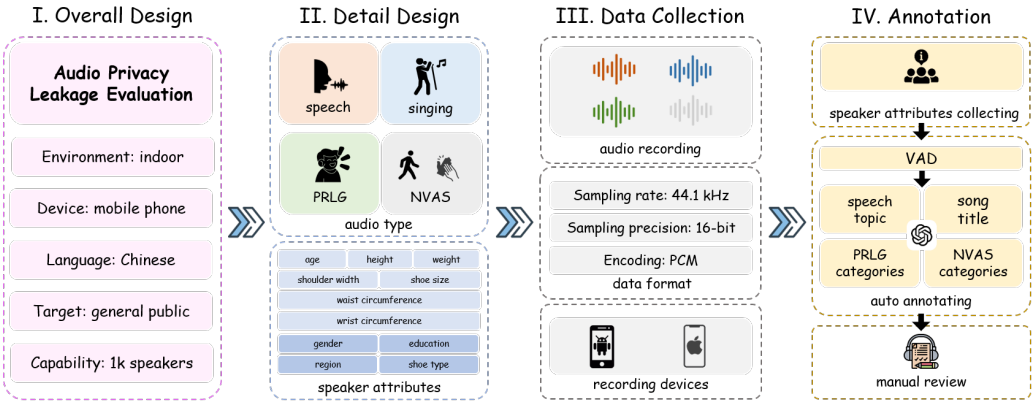


Figure 2: Overview of the AudioPrivacy data construction process, including overall design, detail design, data collection, and annotation.

lects Mandarin and eight regional dialects from 27,237 speakers across various regions of China, with additional annotations for speaker location and dialect type. MIX6 (Brandschain et al., 2010), a large-scale ASR-focused dataset, comprises 15,863 hours of audio recorded across 14 channels, including interviews, transcribed readings, and conversational phone speech.

However, ASR-oriented datasets typically offer sparse attribute annotations, which minimize the differences between individual speakers. These datasets are designed to focus on common speech features, not personalized speaker traits. As a result, their value in analyzing unique speaker attributes is limited.

2.2 Datasets for Speaker Recognition

Datasets designed for Speaker Recognition (SR) tasks can also be leveraged for speaker profiling.

The well-known VoxCeleb series (Nagrani et al., 2019; Chung et al., 2018) includes over 10,000 voice segments from 1,251 celebrities in VoxCeleb1 (Nagrani et al., 2019), with annotations covering age, gender, race, occupation, and accent. VoxCeleb2 (Chung et al., 2018) expands this to 6,000 speakers and over 1,000,000 voice segments. The NIST Speaker Recognition Evaluation (SRE) dataset is another resource that can be applied to speaker profiling.

CNCeleb (Fan et al., 2020), which collected 500 hours of audio from 3,000 celebrities on the Chinese platform BiliBili, covers 11 complex acoustic conditions in various speech scenarios. While these datasets offer a large speaker pool and sufficient audio data, the range of annotated speaker attributes remains limited. HeightCeleb (Kacprzak and Kowalczyk, 2024), built upon VoxCeleb1, adds height information to enhance the attribute variety.

Overall, these datasets primarily focus on identity differentiation and provide shallow attribute annotations. Most data is sourced from celebrity interviews and media appearances, which may not reflect the everyday conditions of ordinary individuals. This discrepancy can pose challenges for privacy-related tasks involving speaker data.

2.3 Datasets for Speaker Profiling

Datasets specifically designed for speaker profiling are typically tailored to specialized domains.

NISP (Kalluri et al., 2021) designed 5 physical attributes and English accent information for describing speakers. VCTK-RVA (Sheng et al., 2025) is an extension of the VCTK (Yamagishi et al., 2019) corpus, highlighting is the addition of expert pairwise annotations for 18 fine-grained voice attributes, such as "Bright" and "Coarse", with some attributes also distinguishing by gender. SPGISpeech 2.0 (Grossman et al., 2025) is a multi-speaker transcription dataset designed for the financial domain, offering dedicated resources for speaker profiling applications in the financial sector.

These datasets do not provide comprehensive descriptions of speaker attributes and were not specifically designed for speaker privacy, which leads to challenges in adapting them for privacy-related tasks.

3 Dataset Description

3.1 Dataset Introduction

This subsection follows the dataset construction pipeline illustrated in Fig.2 and introduces the

type	content	# utt	# dur (h)	avg (s/u)
SPCH	topic	39027	122.19	11.27
SING	acappella	17381	48.09	9.96
PRLG	'wei'	7954	2.33	1.05
	coughing	7975	2.85	1.28
	laughing	7893	3.02	1.38
	crying	7923	4.75	2.16
	yawning	7810	4.22	1.95
	'ah'	7891	2.49	1.14
	'eh'	7884	2.53	1.16
	sighing	7943	2.94	1.33
	total	63431	25.20	1.43
NVAS	walk	4966	12.51	9.07
	jump	4984	1.86	1.34
	clap	4968	2.08	1.51
	upstairs	3000	7.99	9.59
	downstairs	2991	7.34	8.84
	total	20951	31.84	5.47
Total		140789	227.32	5.81

Table 2: Duration for each type of audio, where SPCH, SING, PRLG, NVAS representing the audio type speech, singing, paralanguage, and no-vocal acoustic signal respectively.

dataset from four aspects.

3.1.1 Overall Design

The dataset presented in this work is designed for speaker privacy leakage risk evaluation. It aims to study and address privacy risk assessment arising from potential audio eavesdropping in everyday indoor environments encountered by the general public. Audio data are collected from 1,000 native Chinese speakers. Each speaker uses their own smartphone for recording, as smartphones are the most commonly used and easily accessible microphone devices in daily life and, consequently, the most likely sources of audio leakage.

3.1.2 Detail Design

Besides speech, the possibility that other types of audio may leak SAP cannot be excluded. Therefore, for each speaker, we collect four categories of parallel audio data as described in Tab. 2: SPCH, SING, PRLG, and NVAS.

Speech data are recorded by prompting speakers to talk about topics they are interested in or familiar with, without including the voice of the prompter. Singing data consist of a cappella recordings of songs that the speakers are familiar with or enjoy, without instrumental accompaniment. Paralinguistic audio includes eight basic vocal expressions: "hello", coughing, laughter, crying, yawning, "ah", "uh" and sighing. To ensure diversity in paralinguistic expression, each sound is performed five times using different expressive styles. NVAS comprise hand clapping and footsteps. For hand clapping, speakers are instructed to clap consecutively from one to five times, while footstep recordings include walking and jumping on flat ground, as well as ascending and descending stairs under different motion conditions.

split	#spk	# utt	durations (h)	avg dur (s/u)
train	800	112736	181.88	5.81
dev	100	14070	22.82	5.84
eval	100	13983	22.62	5.82

Table 3: Details of each dataset split

To cover a broader spectrum of potential SAP leakage, the dataset annotates 11 speaker attributes. In addition to age and gender, which are commonly included in existing datasets, we also collect height, weight, shoulder width, waist circumference, wrist circumference, province, education level, shoe type, and shoe size. These speaker attributes are all potentially correlated with acoustic signals but remain insufficiently explored in prior studies.

3.1.3 Data Collection

AudioPrivacy is recorded using the speakers' own Android and iOS smartphones and comprises 140,789 utterance-level audio samples, with a total effective duration of 227.32 hours. To preserve fine-grained acoustic details for future research, all recordings are captured at a sampling rate of 44.1 kHz with 16-bit precision using PCM encoding. During data collection, speaker information is carefully curated to maintain a relatively balanced distribution across different SAP dimensions.

3.1.4 Annotation

To enable broader applicability, AudioPrivacy provides content annotations not only for PRLG and NVAS audio, but also for speech and singing audio through multimodal large language models. In the 'extra_info' field of the audio metadata, speech recordings are annotated with topic-related keywords, while singing recordings are labeled with the corresponding song titles. The semantic information contained in these audio samples may further increase the risk of SAP leakage.

3.2 Dataset Splits

The dataset partitions the 1,000 speakers into training, validation, and evaluation sets with a ratio of 8:1:1, as shown in Tab.3. Given that speaker

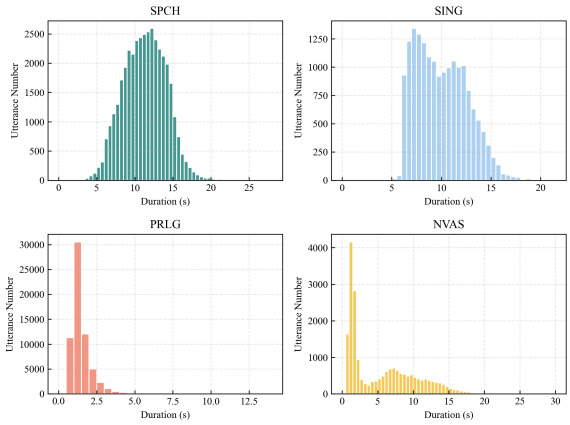


Figure 3: The bar charts of utterance-level duration distribution statistics of 4 types of audio

attributes span multiple dimensions and include both discrete labels and continuous values, and that their distributions need to be balanced, continuous attributes are first discretized into intervals. Subsequently, the speakers are split using the MultilabelStratifiedShuffleSplit [2]tool from scikit-learn, ensuring that the attribute distributions across the three subsets remain approximately independent and identically distributed.

3.3 Statistics

Fig. 3 presents the utterance-level duration distributions for the four audio categories. Speech durations approximately follow a normal distribution, with an average length of around 13 s. Singing exhibits a bimodal distribution, with two prominent peaks near 7.5 s and 12 s, which is likely caused by the segmentation constraint enforcing a minimum utterance length of over 5 s. PRLG audio durations are generally short and concentrated around 1.5 s. In contrast, NVAS also display a bimodal pattern: short-duration peaks correspond to transient events such as clapping and jumping, while longer durations arise from walking and stair-climbing recordings, which require capturing the full motion trajectory from near to far. The speaker distribution statistics for each attribute are detailed in Sec.A.1.

4 Tasks and Baselines

This section experimentally validates the effectiveness of AudioPrivacy for audio privacy leakage evaluation. To align with common experimental settings in audio research, all experiments are conducted after downsampling AudioPrivacy to 16 kHz. We design evaluation experiments to assess the privacy leakage risks induced by two tasks: SV and SP.

[2]https://github.com/trent-b/iterative-stratification

Baselines	batch_size	trainable parameters (M)
ECAPA-TDNN	128	20.9
ResNet-TDNN	32	15.7
WavLM-ECAPA	128	20.9

Table 4: The baselines' configuration for the SV task on AudioPrivacy

4.1 Speaker Verification

SV task aims to evaluate whether different types of audio can be associated with speaker identity, thereby leading to identity leakage. Since NVAS are not produced by the vocal tract and are therefore inherently independent of speaker identity, they are excluded from the SV experiments.

4.1.1 Data Processing

For the SV task on AudioPrivacy, speakers from the original training and validation sets are merged and re-partitioned to construct new training and evaluation sets. For each audio type, utterances corresponding to the same speaker are split into training and test subsets with a 9:1 ratio. Speakers from the validation set are reserved for the verification stage, where enrollment and trial sets are constructed as follows: for each evaluation speaker, the first utterance is used for enrollment, and all remaining utterances are used as trials. The maximum duration of each utterance is limited to 3 seconds.

4.1.2 Baseline Configuration

Experiments are conducted using three network architectures for speaker classification training: ECAPA-TDNN [3](Desplanques et al., 2020), ResNet-TDNN [4](Villalba et al., 2020), and ECAPA-TDNN with WavLM [5](Chen et al., 2022) as the feature extractor (WavLM-ECAPA). Speaker verification inference is then performed on enrollment–trial pairs, with Equal Error Rate (EER) and Minimum Detection Cost Function (minDCF) used as evaluation metrics. Detailed model configurations are provided in Tab. 4. The numbers of trainable parameters for ECAPA-TDNN, ResNet, and WavLM-ECAPA are 20.9M, 15.7M, and 20.9M,

Audio Type	Model	EER (%)	minDCF
SPCH	ECAPA-TDNN	0.82	0.0007
	ResNet-TDNN	0.26	0.0002
	WavLM-ECAPA	0.97	0.0010
SING	ECAPA-TDNN	2.29	0.0018
	ResNet-TDNN	1.64	0.0011
	WavLM-ECAPA	2.21	0.0019
PRLG	ECAPA-TDNN	33.81	0.0099
	ResNet-TDNN	32.09	0.0100
	WavLM-ECAPA	20.65	0.0097

Table 5: The performance of three baselines evaluating on AudioPrivacy for SV

respectively.

Following the SpeechBrain (Ravanelli et al., 2021) recipe, ECAPA-TDNN and ResNet-TDNN are initialized from pretrained models. For WavLM-ECAPA, the WavLM backbone is frozen as a feature extractor, and an ECAPA-TDNN backend is trained from scratch. The extracted WavLM-Large features are obtained by averaging the summed representations across all transformer layers.

Softmax-AAM (Deng et al., 2019) is adopted as the training loss, and CyclicLR is used as the learning rate scheduler, with the learning rate varying from 1×10^{-8} to 1×10^{-3} and the step size is 65,000. Due to differences in data scale, models are trained for 20, 30, and 50 epochs on the SPCH, PRLG, and SING audio subsets, respectively.

4.1.3 Results of Speaker Verification

Based on the evaluation results Tab.5 on the SV task, we draw the following conclusions: 1) The ability of different audio types to convey speaker identity follows the order: SPCH > SING > PRLG. This trend may be partly attributed to differences in corpus duration, as this ordering coincides with decreasing average utterance length. 2) Despite having an average utterance duration of less than 1.5 s, PRLG still achieves an EER of 20.65%, indicating that paralinguistic information remains susceptible to identity leakage when intercepted. 3) Audio types with richer semantic and phonetic content, such as SPCH and SING, are more likely to cause identity leakage. Although PRLG carries partial identity cues, it is insufficient on its own to pose a strong threat to speaker identity.

[5] https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

[5] https://huggingface.co/speechbrain/spkrec-resnet-voxceleb

[5] https://huggingface.co/microsoft/wavlm-large

4.2 Speaker Profiling

SP tasks are conducted under three different paradigms: traditional deep learning (DL) models, pretrained MM LLMs, and the fine-tuned MM LLM, where the traditional DL and LLM pipelines are implemented with SpeechBrain (Ravanelli et al., 2021) and Swift (Zhao et al., 2024) frameworks respectively.

Differences in task types and value ranges across speaker attributes make it difficult to adopt a unified risk evaluation metric. To enable cross-attribute and cross-range assessment of audio privacy leakage, this paper proposes Relative Attribute Leakage Gain (RALG), a metric designed to quantify the additional privacy risk introduced by a target model compared to random guessing in the absence of audio input. Using RALG, SAP risks can be uniformly mapped to the ([0, 1)) range for direct comparison. The RALG metric is defined as follows:

$$\text{RALG} = \begin{cases} 1 - \dfrac{1 - F1_{\text{test}}}{1 - F1_{\text{maj}}}, & if \text{ classification}, \\ 1 - \dfrac{MAE_{\text{test}}}{MAE_{\text{med}}}, & if \text{ regression}. \end{cases}$$

Here, $F1_{\text{maj}}$ represents the Macro F1 score achieved by predicting the majority class once the set of class labels is known, while MAE_{med} denotes the mean absolute error obtained by always predicting the median value given only the lower and upper bounds of the ground-truth values.

In the experimental results, tables are visualized using color gradients. Higher RALG values, corresponding to a greater privacy leakage risk, are highlighted with deeper shades of red to facilitate intuitive comparison across SAP attributes.

4.2.1 Traditional Deep Learning Baselines

Traditional deep learning approaches typically learn a mapping from a single audio category to a single speaker attribute, treating discrete attributes and continuous attributes as classification and regression tasks, respectively. This paradigm requires training separate predictive models for each specific audio–attribute pair. While straightforward, this approach poses a stronger potential threat in terms of SAP leakage, as each model is explicitly optimized to infer a particular speaker attribute from the audio signal.

The experiments adopt two models: Fbank + ECAPA-TDNN [6], representing the traditional handcrafted feature plus DNN paradigm, and WavLM

	Fbank + ECAPA-TDNN				WavLM + linears			
	SPCH	SING	PRLG	NVAS	SPCH	SING	PRLG	NVAS
age	32.00	26.25	11.10	14.74	24.95	25.35	11.29	14.29
height	37.30	35.93	39.39	22.63	40.93	40.47	33.75	24.21
weight	43.93	35.08	44.81	30.31	34.53	31.73	29.65	38.67
shoulder width	36.90	28.11	28.52	31.74	23.08	18.00	34.26	30.08
waist circumference	40.15	45.02	35.78	37.55	45.53	38.71	44.53	45.34
wrist circumference	44.31	40.79	38.08	49.79	52.38	50.94	52.57	53.99
shoe size	21.13	33.93	34.60	4.13	24.00	29.46	27.02	0.00
gender	92.61	85.58	77.61	31.17	93.26	91.17	74.71	23.56
region	27.78	20.82	22.44	32.65	29.37	15.54	29.42	28.24
education	8.51	0.00	6.10	15.20	0.00	0.00	0.00	0.00
shoe type	14.35	15.87	13.43	21.00	13.40	9.69	10.78	12.13

Table 6: RALG evaluation of Traditional Deep Learning Baselines across audio types and SAPs.

	MiMo-Audio-7B-Base				Qwen2.5-Omni-3B				Qwen2.5-Omni-7B			
	SPCH	SING	PRLG	NVAS	SPCH	SING	PRLG	NVAS	SPCH	SING	PRLG	NVAS
age	5.51	5.32	6.10	6.03	0.00	0.00	0.00	0.00	21.30	3.64	0.00	0.00
height	28.69	28.64	30.14	29.95	32.03	31.66	30.22	29.40	40.56	29.70	34.01	29.40
weight	13.76	13.41	14.29	14.37	13.17	23.42	5.33	0.00	0.00	2.75	34.58	18.92
shoulder width	35.67	35.58	35.11	35.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	17.37
waist circumference	10.44	11.55	10.38	10.24	0.00	0.00	0.00	0.00	0.91	0.00	1.84	0.00
wrist circumference	31.21	31.69	31.64	31.30	19.07	32.88	30.67	10.98	35.37	28.04	33.33	0.00
shoe size	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
gender	0.00	0.00	0.00	0.00	86.23	88.15	69.21	25.78	42.81	44.08	37.40	0.00
region	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.39	0.69	1.28	1.40
education	0.00	0.00	0.00	0.00	0.00	0.00	2.05	0.79	0.00	0.00	0.00	0.00
shoe type	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	4.12	3.65

Table 7: RALG evaluation of pre-trained MM LLM across audio types and SAPs

[7], representing the end-to-end self-supervised pre-training paradigm. To fully expose potential privacy leakage risks, both ECAPA-TDNN and WavLM are initialized from pretrained checkpoints on large-scale speaker datasets. Specifically, ECAPA-TDNN is initialized from a version pretrained on CN-Celeb, while WavLM uses a checkpoint pretrained for speaker verification on VoxCeleb. For each model, separate downstream heads are constructed for classification and regression tasks, each consisting of four linear layers.

For each audio–attribute pair, ECAPA-TDNN and WavLM are trained with learning rates of 1×10^{-3} and 1×10^{-4}, respectively. Models are trained for 20 epochs on the training set, and the checkpoint achieving the best performance on the validation set is selected for evaluation on the evaluation set. The objective of these traditional deep learning experiments is to investigate the feasibility of inferring SAP from acoustic information; therefore, the maximum duration of audio input is limited to 3 seconds. The experimental results lead to the following valuable conclusions:

- SAPs such as height, weight, waist circumference, and wrist circumference exhibit a high level of leakage, with the RALG risk exceeding 30% for nearly all audio categories.
- For SAPs like age, height, waist circumference, wrist circumference, and shoe type, more information can be extracted from NVAS audio than from PRLG audio.
- Vocal acoustic features are more effective for identifying physical attributes, whereas NVAS audio shows a stronger correlation with shoe type and waist circumference. This suggests that the acoustic information in NVAS is more closely related to the material and structure of objects causing vibrations.

Notably, SPCH and SING outperform NVAS in shoe size prediction, which seems counter-intuitive. This suggests that speech serves as a reliable biological proxy; the vocal tract's physiological correlation with the skeletal frame provides stable latent features for inferring physical dimensions. In contrast, NVAS (especially footsteps) represents extrinsic impact sounds confounded by material-related noise (e.g., sole and flooring types), which obscures intrinsic biological traits.

4.2.2 Pre-trained LLM Baselines

MM LLMs have already acquired rich acoustic knowledge through large-scale audio pretraining and can unify multi-attribute prediction via textual

[7] https://huggingface.co/LanceaKing/spkrec-ecapa-cnceleb

[7] https://huggingface.co/microsoft/wavlm-base-sv

	Qwen2.5-Omni-3B			
	SPCH	SING	PRLG	NVAS
age	**37.78**	**33.63**	**13.49**	10.51
height	**53.54**	**49.74**	**42.59**	13.27
weight	**44.56**	**42.86**	36.13	22.90
shoulder width	24.69	25.60	28.46	24.39
waist circumference	43.81	43.77	39.34	29.61
wrist circumference	49.31	48.61	47.03	41.34
shoe size	**47.76**	**45.99**	**39.38**	0.00
gender	**94.90**	**94.68**	**87.56**	**31.54**
region	**41.29**	**38.40**	**30.62**	**34.48**
education	**28.39**	**27.71**	**13.85**	**16.49**
shoe type	**20.67**	**19.59**	**17.67**	14.08

Table 8: RALG evaluation of the SFT Qwen2.5-Omni-3B model across four audio categories and eleven SAPs.

outputs. By analyzing their inference results on the evaluation set of AudioPrivacy, we evaluate whether state-of-the-art MM LLMs—representing the current frontier of artificial intelligence—have developed sufficiently strong speaker profiling capabilities and, consequently, whether they pose a tangible threat to audio privacy.

The paper evaluates three multimodal LLMs on the AudioPrivacy evaluation set: MiMo-Audio-7B-Base [8], Qwen2.5-Omni-3B [9], and Qwen2.5-Omni-7B [10]. To ensure that the MM LLMs adequately understand the audio inputs and produce outputs in a standardized format, task-specific user prompts are carefully designed for different audio types and SAP inference tasks. Detailed prompt templates are provided in Sec.A.2.1.

As expected, even without task-specific training, pretrained MM LLMs exhibit non-negligible leakage risks for body-related SAPs, while their inference performance for shoe size is the weakest. The results shown on Tab.7 suggest that LLMs are already capable of perceiving certain aspects of speakers physique based on the rich acoustic and semantic knowledge acquired during pretraining. Nevertheless, the extent to which MM LLMs pose a threat to SAP remains an open question and requires further experimental investigation.

4.2.3 Fine-tuned LLM

Although pretrained MM LLMs do not appear to exhibit strong SAP inference capabilities, it is still necessary to investigate whether their performance can be significantly improved after supervised fine-tuning (SFT). In comparison with all aforementioned models, the highest RALG values achieved by SFT LLMs are highlighted in bold font, emphasizing the potential risks that fine-tuned LLMs may pose to speaker attribute privacy.

Under limited computational resources, we select the smallest model among the evaluated LLMs, Qwen2.5-Omni-3B, and perform supervised fine-tuning for only three epochs. During SFT, bottleneck adapters with a hidden dimension of 256 and an adapter length of 128 are inserted into the q,k,v, and out projections of the speech encoder's self-attention layers. In addition, LoRA adapters with rank $r = 8$, scaling factor $alpha = 32$, and dropout rate 0.05 are applied to the q, k, v, o, gate, up, and down projection parameters within the self-attention modules of the thinker LLM.

To further reduce computational cost, both training and inference are formulated such that a single audio input corresponds to the prediction of all SAP dimensions simultaneously, thereby avoiding redundant audio modality inputs. Detailed configurations are provided in Sec.A.2.2.

The experimental results in Tab. 8 highlight a significant threat to SAPs posed by MM LLMs, raising serious concerns. Fine-tuning the Qwen2.5-Omni-3B model leads to substantial performance improvements across nearly all SAP inference tasks, outperforming traditional deep learning models in over half of the SAP inferences. This indicates that large models are capable of modeling both audio and multi-dimensional SAPs, creating a more detailed, consistent, and interconnected representation of the speaker's profile. As a result, audio privacy is confronted with an unprecedented level of threat.

5 Conclusion

This paper introduces AudioPrivacy, a large-scale Chinese dataset designed to support systematic evaluation of speaker attribute privacy in realistic audio leakage scenarios. By covering diverse audio types and rich speaker attributes, the dataset enables fine-grained analysis of privacy risks beyond voiceprint-based identity protection. Comprehensive experiments reveal that multiple speaker attributes, including those inferred from non-speech audio, are highly vulnerable to acoustic leakage. The results further indicate that advanced multimodal models can exacerbate such risks under unified evaluation. Overall, this work highlights the urgency of rethinking audio privacy protection in the era of powerful foundation models.

[10] https://huggingface.co/XiaomiMiMo/MiMo-Audio-7B-Base

[10] https://huggingface.co/Qwen/Qwen2.5-Omni-3B

[10] https://huggingface.co/Qwen/Qwen2.5-Omni-7B

Acknowledgements

This work was supported by the National Key R&D Program of China (Grant No. 2022ZD0116307) and the National Natural Science Foundation of China (Grant No. 62271270).

Limitations

Regarding speaker selection, the distribution of speakers across individual provinces may not be perfectly uniform. Given a dataset size of 1,000 speakers, ensuring an even distribution across all 34 provincial-level administrative regions is challenging. As a practical compromise, we instead ensure that speakers are relatively balanced across the seven major geographic regions of China.

Due to limited computational resources and the large scale of the experiments, we were unable to conduct sufficient repeated runs to fully mitigate experimental variance. In addition, the scope of model evaluation in this work remains relatively constrained, and the inclusion of a wider variety of architectures, training strategies, and model scales may provide a more comprehensive characterization of SAP leakage risks across different attributes. In addition, the current analysis treats each audio type in isolation, whereas real-world adversaries may jointly exploit multiple recordings or heterogeneous audio sources from the same speaker. Such multi-instance and cross-type audio fusion could introduce complementary cues and further amplify SAP leakage, warranting systematic investigation in future work.

Ethics Statement

This study was conducted in accordance with strict ethical guidelines to ensure the safety, privacy, and well-being of all participants. All participants were over 18 years of age, and all audio recordings were collected with the informed consent of each individual. Data collection took place in the participants' own living environments, using recording devices provided by the participants themselves. During the recording process, participants were free to choose familiar or preferred topics and songs, without pressure or exposure to sensitive prompts. Each participant received appropriate compensation of 150 RMB (approximately USD 20) for their time and contribution.

To ensure the appropriateness of the audio content and protect participant privacy, all recordings were manually reviewed. Any audio containing personally identifiable information, such as names or contact details, was excluded from the publicly released dataset. The final dataset is fully anonymized, and all metadata or content that could enable re-identification has been removed.

The dataset is distributed exclusively for academic and non-commercial research purposes, and all users must formally agree to the Terms of Access. These terms explicitly prohibit commercial use, participant re-identification, data redistribution, and unethical applications. In particular:

- All recipients are required to delete the relevant data upon a participant's request for removal;
- Researchers must comply with applicable institutional ethical review protocols (e.g., IRB) and ensure that the dataset is not misused for surveillance, profiling, or other harmful purposes;
- Derivative works may not be distributed beyond the research group without explicit permission.

The dataset is released under a non-commercial research license (CC BY-SA-NC 4.0) and must be properly cited in any derivative publications or presentations. The data are provided "as is" without warranty, and the maintainers reserve the right to revoke access in cases of policy violation.

Through these measures, we aim to maximize the dataset's value to the research community while upholding the highest ethical standards of data protection. Such safeguards are essential for responsible research on audio privacy.

Statement on AI Usage

The authors employed AI assistants (e.g., ChatGPT/Claude) solely for language polishing and the refinement of scientific illustrations and plots. All core research components—including the conceptualization, experimental design, data analysis, and research conclusions—were conducted independently by the authors. The authors have meticulously reviewed all outputs and assume full responsibility for the integrity, accuracy, and originality of the final manuscript.

References

Massa Baali, Shuo Han, Syed Abdul Hannan, Purusottam Samal, Karanveer Singh, Soham Deshmukh, Rita

Singh, and Bhiksha Raj. 2025. Colmbo: Speaker language model for descriptive profiling. *arXiv preprint arXiv:2506.09375*.

Tom Bäckström. 2025. Privacy in speech technology. *Proceedings of the IEEE*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Linda Brandschain, David Graff, Chris Cieri, Kevin Walker, Chris Caruso, and Abby Neely. 2010. Mixer 6. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2026–2033.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Peng Cheng and Utz Roedig. 2022. Personal voice assistant security and privacy—a survey. *Proceedings of the IEEE*, 110(4):476–507.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *Interspeech 2018*, pages 2206–2210.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Interspeech 2020*.

Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang. 2020. CN-Celeb: a challenging chinese speaker recognition dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6844–6848. IEEE.

John Garofolo, L Lamel, W Fisher, Jonathan Fiscus, D Pallett, and Nancy Dahlgren. 1993. Darpa timit acoustic-phonetic continuous speech corpus cd-rom TIMIT.

Raymond Grossman, Taejin Park, Kunal Dhawan, Andrew Titus, Sophia Zhi, Yulia Shchadilova, Weiqing Wang, Jagadeesh Balam, and Boris Ginsburg. 2025. Spgispeech 2.0: Transcribed multi-speaker financial audio for speaker-tagged transcription. In *Proc. Interspeech 2025*, pages 4048–4052.

Simon Hanisch, Patricia Arias-Cabarcos, Javier Parra-Arnau, and Thorsten Strufe. 2025. Anonymization techniques for behavioral biometric data: a survey. *ACM Computing Surveys*, 57(11):1–54.

Stanisław Kacprzak and Konrad Kowalczyk. 2024. Heightceleb-an enrichment of voxceleb dataset with speaker height information. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 857–862. IEEE.

Shareef Babu Kalluri, Deepu Vijayasenan, and Sriram Ganapathy. 2020. Automatic speaker profiling from short duration speech data. *Speech Communication*, 121:16–28.

Shareef Babu Kalluri, Deepu Vijayasenan, Sriram Ganapathy, Prashant Krishnan, and 1 others. 2021. Nisp: A multi-lingual multi-accent dataset for speaker profiling. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6953–6957. IEEE.

Wenyu Li, Xiaoqi Jiao, Yi Chang, Guangyan Zhang, and Yiwen Guo. 2025. Audiorole: An audio dataset for character role-playing in large language models. *arXiv preprint arXiv:2509.23435*.

Yuke Lin, Ming Cheng, Fulin Zhang, Yingying Gao, Shilei Zhang, and Ming Li. 2024. Voxblink2: A 100k+ speaker recognition corpus and the open-set speaker-identification benchmark. *Interspeech 2024*.

Xiaoxiao Miao, Ruijie Tao, Chang Zeng, and Xin Wang. 2025. A benchmark for multi-speaker anonymization. *IEEE Transactions on Information Forensics and Security*.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2019. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:100995.

Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, Els Kindt, Héctor Delgado, Massimiliano Todisco, Mohamed Amine Hmani, Aymen Mtibaa, and 1 others. 2019. Preserving privacy in speaker and speech characterisation. *Computer Speech & Language*, 58:441–480.

Authors of SonicSet. 2024. Sonicsim: A customizable simulation platform for speech processing in moving sound source scenarios (referencing sonicset). *arXiv preprint arXiv:2410.01481 (Please use the published version if available)*.

Authors of WildElder. 2025. Wildelder: A chinese elderly speech dataset from the wild with fine-grained manual annotations. *arXiv preprint arXiv:2510.09344 (Please use the published version if available)*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *ICASSP 2015-2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, and 1 others. 2021. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Zhao Ren, Kun Qian, Tanja Schultz, and Björn W Schuller. 2023. An overview of the icassp special session on ai security and privacy in speech and audio processing. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, pages 1–6.

Seyed Omid Sadjadi, Craig Greenberg, Elliot Singer, Lisa Mason, and Douglas Reynolds. 2022. The 2021 nist speaker recognition evaluation. *arXiv preprint arXiv:2204.10242*.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

Zheng-Yan Sheng, Li-Juan Liu, Yang Ai, Jia Pan, and Zhen-Hua Ling. 2025. Voice attribute editing with text prompt. *IEEE Transactions on Audio, Speech and Language Processing*.

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xi Tan, Chuandong Xie, Shuran Zhou, and 1 others. 2021. Kespeech: An open source speech dataset of mandarin and its eight subdialects. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 586–590. IEEE.

Jesús Villalba, Nanxin Chen, David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Jonas Borgstrom, Leibny Paola García-Perera, Fred Richardson, Réda Dehak, and 1 others. 2020. State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations. *Computer Speech & Language*, 60:101026.

Shilong Wu, Hang Chen, and Jun Du. 2025. M3sd: Multi-modal, multi-scenario and multi-language speaker diarization dataset. *arXiv preprint arXiv:2506.14427*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive:(http://web. ku. edu/˜ idea/readings/rainbow. htm)*.

Guoli Zhang, Yuxuan Chen, Nanxin Wang, Yanzhuo Li, Yejia Chen, Ji Liu, Yishu Chen, Shuming Wang, Xiang Chen, and Dong Wang. 2021. Voxpopuli: A large-scale multilingual corpus for speech recognition, summarization, and translation from european parliament interviews. *arXiv preprint arXiv:2101.07185*.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. Swift:a scalable lightweight infrastructure for fine-tuning. *Preprint*, arXiv:2408.05517.

Siqi Zheng, Luyao Cheng, Yafeng Chen, Hui Wang, and Qian Chen. 2023. 3d-speaker: A large-scale multi-device, multi-distance, and multi-dialect corpus for speech representation disentanglement. *Preprint*, arXiv:2306.15354.

A Appendix

A.1 Statistics on Speaker Attribute Distribution

We analyze both numerical and categorical speaker attributes, visualizing their distributions using histograms and bar charts, respectively. In Fig. 4, numerical attributes are reported with height measured in centimeters (cm), age in years, and weight in kilograms (kg). In Fig. 5a, categorical attributes include gender, with two categories (M and F) denoting male and female; region, consisting of seven categories corresponding to different geographic areas of China; education level, divided into five categories—P, J, S, U, and U+, representing primary school, junior high school, senior high school, undergraduate, and postgraduate or above; and shoe type, which includes four categories: Y, L, X, and P, denoting sneakers, sandals/slippers, casual shoes, and leather shoes.

The region attribute information of the speakers is collected by province. The distribution of all speakers across provincial-level administrative regions in AudioPrivacy is shown in Fig.5b

A.2 User Prompt for MM LLM

The user text prompts used for MM LLMs in this study can be divided into two categories: the first category is for pre-trained models, while the second category is for model training and inference.

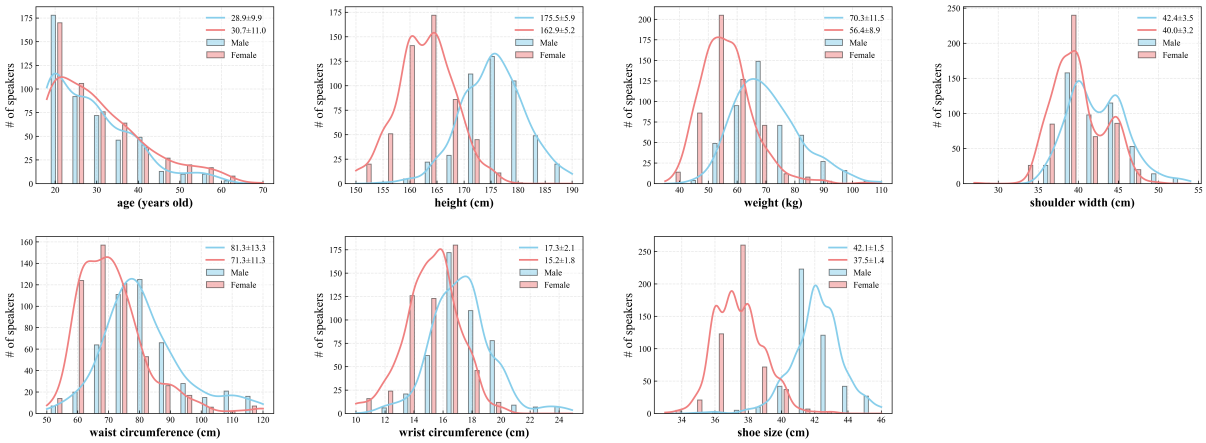
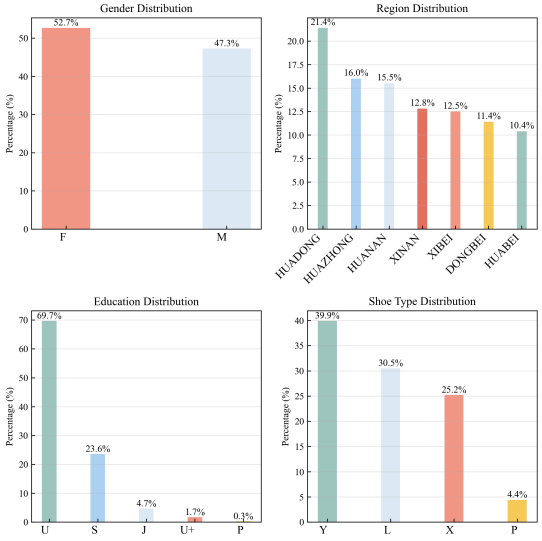
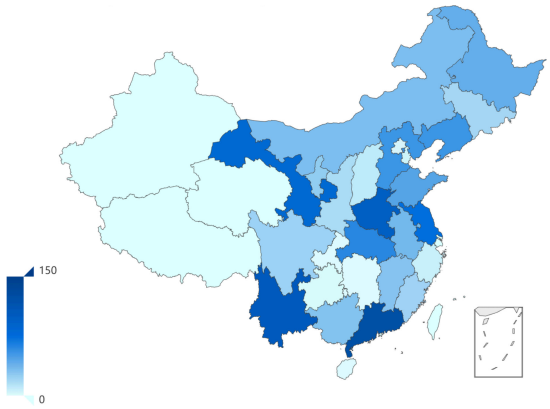


Figure 4: Histogram of speaker numerical attribute distributions, where the value distributions and the corresponding fitted KDE curves are presented separately by gender.



(a) Bar charts of speaker categories attribute distributions.



(b) The spatial distribution map of 1,000 Chinese speakers across province-level administrative areas.

Figure 5: Distribution of categorical speaker attributes and geographical coverage in AudioPrivacy. (a) Statistical breakdown of categorical attributes (Gender, Region, Education, and Shoe Type); (b) Spatial distribution of the 1,000 speakers across provincial-level administrative regions in China.

A.2.1 Pre-trained MM LLM

For the pre-trained MM LLMs, due to their lack of prior exposure to relevant data, there is a high demand for their zero-shot capabilities. In this context, the inference of SAPs by MM LLMs requires the construction of detailed text prompts for guidance. To ensure that the outputs are interpretable and standardized, we designed two different text prompt templates: one for numerical attributes and one for categorical attributes, which are used for audio-to-SAP single-attribute prediction. The following format will be consistently used throughout: <audio> is a placeholder for the audio token, and ${} contains variables executed during Python processing.

Fig.6 presents the prompt template used for categorical attribute inference. In this template, 'spk_SAP_type' refers to the SAP name to be inferred, i.e., the attribute to be predicted; 'output_space' strictly controls the output space, and 'output_explanation' provides an explanation for each of the categorical options.

Fig.7 shows the prompt template used for numerical attribute inference. Here, 'spk_SAP_type' refers to the SAP name, i.e., the attribute to be predicted, while unit sets the unit for the predicted numerical inference.

User Prompt:
<audio>Your task is to infer the speaker's ${spk_SAP_type} from the audio material.

Allowed output space (strict): {{ ${output_space}}}
The outputs represents: ${output_explanation}
Constraints:
- Output must be exactly one of the allowed outputs.
- Just give the most simplified output. Do not output sentences, explanations, or symbols.
- If the output is not one of the allowed outputs, it is considered INVALID.

Now output the classification result:

Figure 6: The user prompt of classification task used for evaluating pre-trained MM LLMs

User Prompt:
<audio>Your task is to infer the speaker's ${spk_SAP_type} from the audio material.

Output a single number (strict) in the unit ${unit} representing the speaker's ${spk_SAP_type}.
Now output the regression result:

Figure 7: The user prompt of regression task used for evaluating pre-trained MM LLMs

A.2.2 SFT MM LLM

For the SFT MM LLM, since multi-dimensional SAPs are output together, the prompt does not need to be tailored for different types and can be pre-written in full. The text prompt and the constructed standardized output format are shown in Fig.8.

User Prompt:
<audio> Infer the speaker's information from the speech audio.
Gender output space: {M, F}, the choices representing: M. male, F. female

Region output space: {HUADONG, HUABEI, HUAZHONG, HUANAN, DONGBEI, XINAN, XIBEI}, the choices representing: HUADONG. Eastern, HUABEI. Northern, HUAZHONG. Central, HUANAN. Southern, DONGBEI. NorthEastern, XINAN. SouthWestern, XIBEI. NorthWestern
Education output space: {P, J, S, U, U+}, the choices representing: P. primary education, J. junior high school, S. senior high school, U. university, U+. post-graduate
Shoe type output space: {X, P, Y, L}, the choices representing: X. Casual shoes, P. Leather shoes, Y. Sneakers, L. Sandals and Slippers
Output a single number (in year) for the age.
Output a single number (in cm) for the height.
Output a single number (in kg) for the weight.
Output a single number (in cm) for the shoulder_width.
Output a single number (in cm) for the waist_circumference.
Output a single number (in cm) for the wrist_circumference.
Output a single number (in EU) for the shoe_size."

Assistance output:
{ "gender": "F", "region": "XIBEI", "education": "U", "shoe_type": "X", "age": "39", "height": "158", "weight": "52", "shoulder_width": "43", "waist_circumference": "69", "wrist_circumference": "16", "shoe_size": "36"}

Figure 8: The data construction for evaluating multi-dimension SAP inference capability of SFT MM LLM