scientific reports

Check for updates

OPEN

Privacy-aware speaker trait and multimodal features relationship analysis in job interviews

Candy Olivia Mawalim[1]✉, Chee Wee Leong[2] & Shogo Okada[1]

As the use of speech data for applications like emotion detection and health profiling grows, so do the privacy risks associated with voice recordings that can reveal sensitive speaker traits. This study investigates voice anonymization methods designed to protect speaker identity while maintaining essential speech characteristics for accurate trait inference, specifically within the context of job interviews. Our experiments show that while anonymization alters several acoustic parameters, the anonymized speech from signal processing-based methods remains suitable for overall trait assessment, with performance comparable to original speech. The phase vocoder-based method, in particular, offers modest privacy gains with an acceptable trade-off in utility, especially in scenarios with minimal attack vectors. In contrast, a neural audio codec-based method altered prosodic features critical for speaker trait estimation, slightly reducing performance in this specific task. Despite this, when carefully configured, this method provides greater privacy and generally preserves utility for speech recognition and quality assessment, even under semi-informed attack scenarios.

Keywords Speaker traits, Human-computer interaction, Privacy protection, Voice anonymization

Automatic speaker trait estimation is an interdisciplinary field, integrating psychology, computer science, and data analytics to infer individual traits from speech. This technology has significant implications for human-computer interaction (HCI) by enabling systems to recognize and adapt to individual characteristics, leading to more personalized and empathetic interactions[1,2]. Automatic speaker trait estimation involves the use of machine learning techniques to infer individual traits from various data sources, such as text, speech, and behavioral patterns[3–7].

In the context of job hiring, where decisions can significantly impact a person's life, the use of automatic speaker trait estimation introduces critical privacy concerns. While this technology could streamline candidate screening, voice data can inadvertently reveal highly sensitive information beyond the traits relevant for a job, such as health conditions, background, and personal vulnerabilities[8,9]. This practice aligns with the concerns outlined in the General Data Protection Regulation (GDPR), which recognizes the risks of profiling and automated decision-making based on personal data[10]. Without robust privacy measures, this data could be misused, leading to algorithmic bias, discrimination, and the violation of a candidate's privacy rights. Therefore, developing and implementing privacy-preserving technologies is not merely a technical consideration but an ethical imperative.

Recent advancements in speaker trait estimation have significantly improved their technical accuracy, but these developments often neglect critical privacy considerations[1,2,6]. As speech-based systems become more prevalent, they pose increasing risks of biometric information leakage[11,12], which raises significant privacy and security concerns that must be addressed to ensure responsible and ethical deployment, particularly in sensitive applications like job hiring.

This study investigates voice anonymization techniques for protecting privacy in automatic speaker trait estimation systems. Motivated by the growing need to balance accuracy with the increasing privacy risks associated with speech data[13], this research contributes to the development of more responsible and ethical AI. By integrating voice anonymization, these systems can leverage speaker trait information while effectively safeguarding individual privacy. Our primary contribution concerns the automated video interview scenario, for which we use a large-scale, real-world dataset collected from an online platform. This approach is grounded in industrial-organizational (I-O) psychology research[14], which provides the theoretical framework for understanding and evaluating personality traits relevant to job performance and hiring decisions. By aligning

[1]Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 923-1292 Nomi, Japan. [2]Educational Testing Service, AI Product and Engineering, Princeton, NJ 08540, USA. ✉email: candylim@jaist.ac.jp

our work with established I-O principles, we ensure our findings on trait assessment are both technically sound and contextually relevant for the hiring process.

Related work

I-O psychology in hiring decisions

The dominant theoretical framework in industrial-organizational (I-O) psychology for understanding and evaluating personality traits relevant to job performance and hiring decisions is the Big Five Personality Traits model (also called Five Factor Model). The Big Five model[15] describes personality traits across five broad dimensions:

- *Openness to experience:* Appreciation for art, emotion, adventure, unusual ideas, curiosity, and a variety of experiences.
- *Conscientiousness:* Tendency to be organized, methodical, disciplined, and goal-oriented.
- *Extraversion:* Enjoyment of social interaction, assertiveness, and excitement.
- *Agreeableness:* Tendency to be compassionate, cooperative, and trusting.
- *Neuroticism:* Tendency to experience negative emotions, such as anxiety, anger, and depression.

Drawing from established I-O psychological frameworks[14], our research on speaker trait estimation for hiring is grounded in several key theories discussed by Kang et al.[16]. *Person-Environment Fit Theory* posits that aligning an individual's personality traits with job requirements leads to increased performance and greater job satisfaction. Furthermore, *Self-Regulation Theory* underscores the importance of traits like conscientiousness and emotional stability, which are critical for an individual's ability to manage their behavior and emotions in the workplace, directly impacting their job performance. Finally, *Social Learning Theory* suggests that personality traits can be shaped by the workplace environment itself, affecting job satisfaction and performance through ongoing social interactions and feedback.

These psychological theories provide the foundational understanding for why and how personality traits influence behavior and outcomes. The field of computational analysis has since sought to quantify these traits, often by analyzing a person's speech. The INTERSPEECH 2012 Speaker Trait Challenge, for example, pioneered the computational analysis of "perceived" speaker traits, including those related to the Big Five model[3,4]. This challenge focuses on predicting personality traits as perceived by human listeners by utilizing acoustic and potentially linguistic features. Prior research has explored various machine learning techniques for speaker trait estimation. For example, support vector machines and hidden Markov models have been used to predict perceived personality traits from speech signals, demonstrating their potential to improve conversational agents[17,18]. The importance of speaker trait estimation has also been investigated in the context of job interviews[5]. This study analyzed speech, prosody, and facial expression analysis with text classification methods to predict candidate personality traits and hiring recommendations. While deep learning has significantly improved the accuracy and efficiency of speaker trait estimation[6], research on privacy-preserving methods for accurate speaker trait estimation from speech remains limited. To the best of our knowledge, this work is among the first to propose such methods.

Privacy-preserving methods for speech processing

Table 1 outlines the key existing privacy-preserving methods for speech applications. Chronologically, these methods have evolved from basic signal-processing techniques to more advanced cryptographic and deep learning-based solutions, reflecting growing regulatory, technical, and societal demands for data protection.

In the context of data sharing, privacy protection can be viewed as an adversarial game. A user seeking to leverage their data for downstream tasks such as automatic speech recognition (ASR) must contend with a

Method	Keypapers	Use case
Homomorphic encryption	Privacy-Preserving Machine Learning for Speech Processing[19]	Secure cloud verification
Voice anonymization	Speaker Anonymization for Personal Information Protection Using Voice Conversion Techniques[20]	Speaker verification
	The VoicePrivacy 2022 Challenge: Progress and Perspectives[21] in Voice Anonymisation	Speaker verification
Attribute removal (age, gender)	Privacy-Oriented Manipulation of Speaker Representations[22]	Speaker verification
Federated learning, differential privacy	Federated Learning with Differential Privacy for End-to-End Speech Recognition[23]	Speech recognition
Differential privacy	Differential privacy for protecting patient data in speech disorder detection using deep learning[24]	Healthcare
Voice anonymization	Addressing challenges in speaker anonymization to maintain utility while ensuring privacy of pathological speech[25]	Healthcare
Cryptographic and speech manipulation	Privacy-preserving Machine Learning for Remote Speech Processing[26]	Healthcare, personal assistance
Adversarial representation	Universal Semantic Disentangled Privacy-preserving Speech Representation Learning[27]	Speech analysis, large language models

Table 1. Representative publications on privacy-preserving speech processing, categorized by use case.

potential attacker. The attacker aims to exploit the shared data or its derivatives to infer sensitive information about the the user[11]. Recent advances in privacy protection for speaker trait estimation leverage anonymization, adversarial learning, and cryptographic techniques to safeguard sensitive information while maintaining data utility.

Initial privacy protection for speech data relied on basic signal processing techniques like spectral warping, pitch shifting, and time-scale modification to mask a speaker's identity. The development of statistical models like Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) introduced new privacy methods based on the statistical models using secure hashing and cryptography. As speaker embeddings (e.g., i-vectors and x-vectors) and voice conversion technologies advanced[28], protecting speaker identity in speaker verification systems became crucial, leading to a high demand for voice anonymization[21]. Today, the focus has shifted toward addressing real-world challenges posed by advanced adversarial attacks, as well as meeting use-case and regulatory requirements.

Voice anonymization (or speaker anonymization) is the focus of this study. These techniques modify speech signals to conceal the speaker's identity while preserving the linguistic content[13]. As voice-based services become more common, the privacy risks associated with sharing speech data have increased, making these methods more important than ever. Signal processing-based approaches, such as McAdam coefficient modification[29] and phase-vocoder-based time-scale modification (PV-TSM)[30], alter acoustic characteristics to obscure speaker-related information. Deep learning-based approaches, such as x-vector-based anonymization[12], phonetic intermediate representations[31], and neural audio codec (NAC)-based anonymization[32], leverage deep neural networks to achieve higher levels of anonymization while striving to retain the utility aspects of speech data. The NAC is an audio codec based on neural networks that compresses audio into discrete digital codes, or tokens, which are then used to reconstruct high-fidelity audio[33]. Further technical details on the specific voice anonymization methods utilized in this work can be found in Section "Voice anonymization methods".

Privacy-preserving speaker trait estimation

Our proposed privacy-preserving personality estimation framework (Fig. 1) leverages voice anonymization to protect user privacy during speaker trait estimation. This research focuses on a data leakage scenario in which an attacker targets the personalized HCI system developed using the original speech. A data leakage scenario refers to the risk that a speaker's private information (e.g., identity or health status) could be inadvertently revealed or extracted from their speech data. While voice anonymization aims to mask individual identities, a determined attacker might still attempt reidentification–trying to correctly link the anonymized voice back to the original speaker's identity–on the basis of subtle cues in the anonymized data. Therefore, we evaluate the verification rate in automatic speaker verification (ASV), which acts as a measure of the attacker's success. A high error in verification rate means the anonymization successfully prevented the attacker from correctly identifying the speaker. We perform this evaluation assuming that the attacker is familiar with the anonymization methods but unaware of the specific parameters used during anonymization.

This framework addresses critical challenges in existing anonymization methods. We argue that for an actual and privacy-preserving speaker trait estimation, same-gender speaker anonymization is essential. Gender is known to be correlated with personality traits[34,35]. For example, research suggests that women often report higher levels of agreeableness, conscientiousness, extraversion, and neuroticism, whereas men tend to score higher on openness to ideas. While such correlations may be biased, we do not want to confound such findings unnecessarily, and leave bias mitigation to another independent effort. Furthermore, we emphasize the need

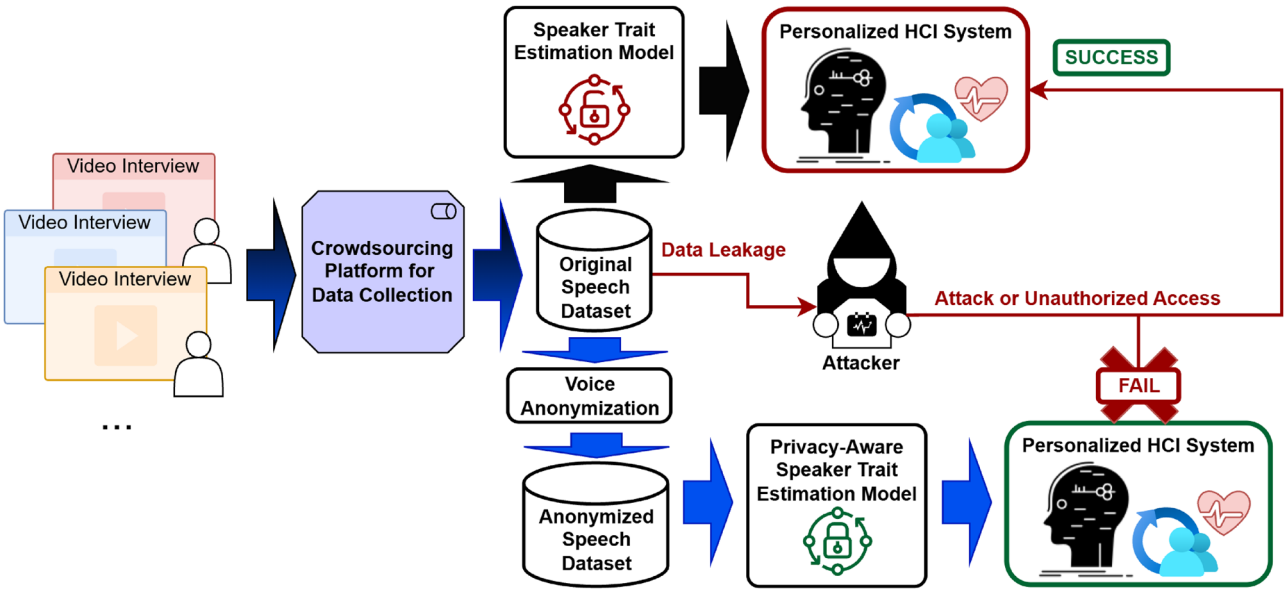

Fig. 1. Scenario for privacy-aware speaker trait estimation from speech signals.

for speaker-level anonymization, rather than utterance-level anonymization. Since individuals answer multiple interview questions, consistent speaker information across all utterances is crucial for accurate personality trait estimation. Utterance-level anonymization could lead to inconsistencies in speaker representation, hindering accurate aggregation of speaker traits.

To evaluate the practicality and limitations of various voice anonymization methods, we investigated both signal processing-based and deep neural network (DNN)-based approaches. While signal processing methods offer speed and ease of implementation, their privacy protection abilities are limited, and they often lack the robustness needed for real-world scenarios featuring noisy or long-duration speech[13]. Existing DNN-based anonymization methods, although generally more robust to noise, present their own challenges. They often do not readily support same-gender anonymization, and adapting these methods for specific applications requires significant further development. For example, preserving gender information in x-vector-based anonymization can compromise privacy[36], which is a critical concern in privacy-aware speaker trait estimation.

Experiments

The objective of this study is to evaluate the effectiveness of various voice anonymization techniques in balancing speaker privacy and speech utility. Our experimental design is structured to systematically test these methods across a range of metrics, including speaker de-identification, intelligibility, and hiring recommendation accuracy. The experiments are conducted using a controlled dataset and a set of predefined anonymization methods, allowing for a reproducible and comparative analysis. This section details the datasets, the specific anonymization methods, and the evaluation setting used in our study.

Dataset

This research utilizes a dataset of recorded online job interviews conducted across the United States collected by Chen at al.[5]. The original dataset comprises 1,891 monologue interview videos, totaling 63 h of content, collected from 260 online workers of diverse ages and racial backgrounds. Each interview followed a structured format featuring questions focused on past work experiences to assess key social competencies. This design aligns with best practices for assessing job-relevant competencies in a systematic manner. Five experts with experience in evaluating written essays and video performances annotated the speaker traits, including a hiring recommendation score and scores on the Big Five personality traits. The statements used for measuring personality traits contained adjectives (e.g., assertive, irresponsible, cooperative) that directly corresponded to the Big Five traits. This measurement approach is similar to the multiple-item measure utilized in prior work[14]. The inter-class correlation values for all annotated scores are higher than 0.75 which provides strong support for the consistent and reliable ground truth data within this dataset. While the raw dataset is derived from video interviews, our current study is strictly focused on voice anonymization.

Voice anonymization methods

For this research, we chose three voice anonymization techniques: McAdam coefficient modification[29], phase-vocoded time-scale modification[30], and neural audio codec (NAC)-based voice anonymization[32]. For brevity, we refer to the McAdam coefficient modification method as MAC, the PV-TSM method as PVT, and the NAC-based method as NAC.

The first two methods–MAC and PVT–represent widely-used signal processing-based approaches that were selected for two primary reasons related to our use case: First, their underlying architecture (linear predictive coding for McAdams, and time-scale modification for PVT) is commonly implemented in speech codecs and audio systems, providing a highly relevant test bed for evaluating utility preservation in real-world communication. Second, their relative ease of implementation and robustness allows them to be applied consistently to the diverse and often noisy audio inputs typical of online job interviews. To preserve gender information, we first estimated the fundamental frequency (F_0) of the original speech. The MAC and PVT methods were then modified with random parameters (the McAdam coefficient ($0.7 \leq \alpha \leq 0.9$) for MAC and the shift parameter ($3 \leq |n| \leq 5$) for PVT), ensuring that the F_0 of the anonymized speech remained within the original speaker's gender-specific F_0 range.

The NAC method was chosen for this study because its architecture is similar to those currently implemented in modern speech communication and audio coding systems, offering a highly relevant design for evaluating real-world utility. Unlike many competing deep learning anonymization techniques, the NAC method operates independently of Automatic Speech Recognition (ASR) systems. This is a crucial practical advantage because ASR is often unreliable when faced with the noisy, unstructured audio input common in real-world online job interviews. By avoiding reliance on transcripts or ASR, the NAC method is inherently more robust, making its performance degradation (or lack thereof) a more reliable measure of the true utility-privacy trade-off for downstream tasks like trait estimation.

Furthermore, the NAC method's architecture is similar to VALL-E[37], a model well-known for generating high-quality synthetic speech when integrated with language models. To meet the specific needs of our speaker trait estimation task, we adapted the NAC method to handle longer utterances, perform anonymization at the speaker level, and preserve gender information. This approach directly addresses the limitations noted in previous work by Panariello et al.[32]. Additionally, other deep learning methods that require transcripts were not feasible for our study, as accurate transcripts were unavailable and ASR outputs in noisy conditions are not sufficiently reliable.

The NAC anonymization was implemented following the B4 recipe provided by VPC 2024[13]. The NAC model used is based on the open-source Bark system[38] (a transformer-based text-to-speech model) and uses Encodec[33] as its core NAC. We adapted it to our speaker trait analysis use case by first modifying the mapping to operate at the speaker level for consistent identity transformation; specifically, we set proxy gender based on gender labels,

mapping male to Bark English speaker ID 6 and female to speaker ID 9. Secondly, to overcome the inherent limitation of Bark generation–which is optimized for short utterances (13–15 seconds) and can exhibit output variations from its internal coarse and fine transformers even with the same proxy speaker selected–we ensured the preservation of the targeted anonymized speaker characteristics across the longer input utterances typical of a job interview by iteratively applying the anonymization process along the entire input length, using identical parameters for each segment to maintain a single, consistent anonymous voice.

Evaluation setting

We evaluated our approach on two downstream tasks: *voice anonymization* and *speaker trait estimation*. For the latter, our study focused on personality traits and hiring recommendations within a job interview scenario.

Voice anonymization

Our evaluation of voice anonymization considered both its privacy and utility aspects, generally following the protocol of the Voice Privacy Challenge (VPC)[13]. The setup for the evaluation is as follows:

Privacy Evaluation: We evaluated privacy using an Automatic Speaker Verification (ASV) system built upon an ECAPA-TDNN model[39]. The Equal Error Rate (EER) served as our primary ASV metric. A lower EER indicates easier speaker verification, reflecting the difficulty of distinguishing between speakers. For this evaluation, we adopted the three attack scenarios defined in the VPC: (a) *Ignorant:* Original enrollment; ASV trained on original data, (b) *Lazy-informed:* Anonymized enrollment; ASV trained on original data, and (c) *Semi-informed:* Anonymized enrollment; ASV trained on anonymized data. While both 'informed' scenarios use anonymized enrollment, they differ in the attacker's knowledge: the Semi-informed attacker knows the anonymization algorithm and retrains the ASV system accordingly, whereas the Lazy-informed attacker uses an ASV system trained on original speech. Conversely, for semi-informed attacks, we used utterance-level anonymized data for training the ASV system. For data partitioning, we followed different approaches based on the dataset:

- *LibriSpeech Dataset:* We used the equivalent partitioning defined in the VPC.
- *Job Interview Dataset:* To reduce computational cost while maintaining performance comparable to the method described in[5], the speech signals were pad to one minute and resampled to 16 kHz. This dataset was divided into non-overlapping training/validation and enrollment/trials sets. Each speaker in this dataset contributed approximately five to eight utterances, each about two minutes in duration. These segments were split 50% for training/validation (80/20 split) and 50% for enrollment/trials (25/75 split). These utterances were then segmented into 20-s clips for ASV evaluation. To optimize the evaluation, we selected approximately 10% random clips per speaker for enrollment and 30% random clips per speaker for trials.

Table 2 provides a comparison of the data used for the ASV evaluation. To manage the computational cost, we utilized only representative clips of approximately 20 seconds in length from the job interview dataset. Using longer utterances for evaluation is crucial, as ASV systems typically yield lower scores with longer speech segments[40]. This approach helps to minimize the possibility of overestimating privacy preservation in a real-world job interview scenario.

Utility Evaluation: We considered two downstream tasks: automatic speech recognition (ASR) and speech quality assessment. For ASR, we calculated the word error rate (WER) via the Whisper model (whisper-large-v3) to transcribe both the original and anonymized speech signals. The WER was then computed by comparing the anonymized transcript (hypothesis) against the original transcript (reference). Subsequently, speech quality was assessed via DNSMOS[41], yielding mean opinion scores (MOS) for the signal (SIG), background (BAK), and overall audio quality (OVR). Higher DNSMOS scores (0–5) indicate better perceived speech quality. DNSMOS is often used to compare noise suppression and speech enhancement algorithms.

Table 2 also compares the data quality of the original LibriSpeech test-clean (Libri test) and job interview datasets using the full datasets. The results highlight a significant disparity in speech quality; the job interview data yielded a considerably lower average MOS (MOS $\approx$ 2.811) compared to the Libri test set (MOS $\approx$ 3.272). This outcome is anticipated, given the uncontrolled and varied acoustic environments inherent in real-world recordings. By comparing the performance of voice anonymization techniques across these two datasets, we can assess the effectiveness of existing methods in preserving privacy under different conditions.

Speaker trait estimation

For speaker trait estimation, we focused on traits that influence hiring recommendations. We conducted an experiment using feature sets that impact the prediction of a hiring recommendation label. These features

Dataset	Average MOS (OVR)	Duration/clip	Subset	#Speakers (F/M)	#Clips
Libri test	3.273 ± 0.198	Mostly < 10s	Enrollment	29 (16/13)	438
			Trial	40 (20/20)	1496
Job interview	2.811 ± 0.325	Approximately 20s	Enrollment	195 (104/91)	1475
			Trial	260 (147/113)	3122

Table 2. Dataset comparison for ASV evaluation: LibriSpeech test-clean (Libri test) vs. job interviews (F and M denote the number of female and male speakers, respectively).

Data	Libri test[42]			Job interview[5]		
Original	4.59			0.49		
Anon. Method	Attack scenario			Attack scenario		
	Ignorant	Lazy-informed	Semi-informed	Ignorant	Lazy-informed	Semi-informed
MAC	21.96	11.81	4.30	10.42	5.43	0.99
PVT	38.15	43.62	12.90	30.54	41.38	17.28
NAC	49.27	49.08	30.27	46.80	42.96	35.96

Table 3. Overall evaluation of voice anonymization methods using ASV in EER (%). An increase in EER after anonymization indicates improved privacy.

System	Metric	Original	MAC	PVT	NAC
ASR	WER (%) (↓)	0.00	5.64 ± 0.11	8.24 ± 4.24	18.85 ± 11.57
DNSMOS	SIG (↑)	3.29 ± 0.28	2.92 ± 0.51	3.13 ± 0.28	3.17 ± 0.40
	BAK (↑)	3.54 ± 0.40	3.17 ± 0.56	3.38 ± 0.40	3.70 ± 0.47
	OVR (↑)	2.81 ± 0.32	2.43 ± 0.44	2.61 ± 0.30	2.79 ± 0.41

Table 4. Overall evaluation of voice anonymization methods using ASR, and DNSMOS. Arrows indicate the direction of improvement: ↑ (higher is better), ↓ (lower is better).

include linguistic information, which we represented using a Bag-of-Words (BoW) model, as it has performed well in prior work[5]. We combined these with a set of paralinguistic features proposed as a baseline in the Interspeech 2012 Speaker Trait Feature Set (IS2012-Trait). We then analyzed the Big Five traits that are highly correlated with hiring recommendations. Our analysis included low-level descriptors derived from the spectral and temporal characteristics of the raw audio signals, e.g., F0 trajectory, jitter, and shimmer. Additionally, we analyzed loudness, rhythm, and timbral features to understand how they affect the speaker traits perception.

To evaluate the speaker trait estimation performance, we used 5-fold cross-validation, stratified by speaker. We incorporated several simple yet effective classifiers that are commonly used in trait estimation, including the Linear Support Vector Classifier (LinearSVC), Random Forest (RF), Light Gradient-Boosting Machine (LightGBM), and eXtreme Gradient Boosting (XGBoost). For all classifiers, we used default hyperparameter settings from their respective Python libraries (e.g., scikit-learn for LinearSVC and RF) and optimized only essential parameters (such as the number of estimators for boosting/forest models) based on performance on the validation set.

Results and discussion

Tables 3 and 4 provide an overview of the privacy-utility trade-off for our voice anonymization methods. Table 3 reports ASV performance to quantify privacy, while Table 4 evaluates utility through ASR-based recognition accuracy and DNSMOS-based speech quality.

Privacy and basic utility analysis

When it comes to privacy, measured by ASV, the methods offer protection in the following order, from least to most effective: *MAC*, *PVT*, and *NAC*. Interestingly, the voice anonymization methods generally provide a higher degree of privacy for the clean Libri test data than for the job interview dataset. This is reflected in the higher EER observed for the Libri test data in the ASV evaluation.

The *MAC* method offers limited protection, achieving an EER of less than 20% in most cases. Critically, in the job interview dataset under a semi-informed attack, the *MAC* EER drops to nearly 0%, indicating virtually no privacy. This presents a significant limitation for data sharing, as the original speaker could be easily traced from the anonymized speech. *PVT* provides medium to high protection, with an EER ranging from 12 to 42%. *NAC* performs best, with its EER increasing to over 30% even in a semi-informed attack scenario.

For our utility analysis, we focused exclusively on the job interview dataset. While the clean Libri test dataset introduces minimal distortion to anonymized speech–a topic covered extensively in the VPC[13,21]–our primary goal was to understand how these anonymization methods perform with a range of noisy, real-world data. The results of the ASR evaluation shown in Table 4 revealed that both *MAC* and *PVT* methods introduce minimal distortion, with WER below 10%. However, *NAC* demonstrated significantly lower accuracy in downstream speech recognition, with a WER approaching 20%. This result contrasts with previous research on the VPC, likely because prior work predominantly used clean datasets. Since the components for our *NAC*-based anonymization were trained on clean speech, they are highly sensitive to the noise present in the job interview dataset. A WER approaching or exceeding 20% is often considered a significant threshold, suggesting that the *NAC* output, in its current state, may not be suitable for real-world deployment where accurate transcription is essential.

As shown in Fig. 2c, the ASR evaluation results for *NAC* correlate with the Mean Opinion Score (MOS) quality bins. Utterances with a MOS less than 2 (poor quality) yielded a median WER of approximately 0.5. In

contrast, recordings with a MOS greater than 3 showed minimal distortion and, consequently, a lower WER. A common issue with ASR on noisy speech, *NAC*-based transcriptions often contained repetitions that were not present in the perceived speech. The simpler modifications used by *MAC* and *PVT* lead to speech that is more similar to the original (average WER of all DNSMOS (OVR) bins is < 0.1). In contrast, *NAC* employs a more sophisticated approach involving semantic, acoustic, and speaker representations, which may contribute to its reduced performance with noisy data.

In terms of DNSMOS, another utility metric, *NAC* provided the best overall quality across SIG, BAK, and OVR, nearly matching that of the original speech. Notably, the average BAK score for *NAC*-anonymized speech even exceeded that of the original speech. This improvement can be explained by noise reduction: when the input signal contained low-energy stationary noise, *NAC* sometimes suppressed it, resulting in a higher BAK score. *PVT* reduced the OVR by approximately 0.2, and *MAC* reduced it by approximately 0.4.

Speaker trait estimation analysis

For speaker trait estimation, we conducted a multistage evaluation to assess the traits most relevant to hiring decisions. We focused our analysis on the Big Five traits. Numerous studies have established a strong link between these traits and job performance across various fields. For instance, Conscientiousness has been consistently identified as the single best predictor of job performance across a wide range of occupations. Additionally, Extraversion is often linked to success in roles that require strong interpersonal skills, such as sales and management.

While many factors influence the outcome of a job interview, in our analysis of a job interview dataset, we found a distinct order of correlation between these traits and hiring recommendations. The most correlated trait was Extroversion (Ex), followed by Conscientiousness (Co), Agreeableness (Ag), Openness (Op), and lastly, Neuroticism (Ne). This finding aligns with the psychological literature, which suggests that extraversion and conscientiousness are influential in professional contexts.

Research in social psychology emphasizes that listeners instinctively and unconsciously attribute personality traits to a speaker upon first hearing their voice. While these attributions may not always be accurate, they are critical as they influence social and professional behavior. Specifically, prosodic features–such as those related to speaking rate and engagement–have been shown to strongly influence perceived personality traits and interviewee performance ratings[43].

In a job interview, the paralinguistic delivery (the how we speak) often carries significant weight alongside the linguistic content (the what we say). A candidate's delivery, if perceived negatively, can lower their hiring recommendation despite a factual answer. Therefore, our study aimed to assess how voice anonymization, by altering acoustic perception, affects the ability to accurately estimate these crucial traits.

To address concerns regarding reproducibility, our trait estimation experiment followed a structured pipeline. We began with Feature Extraction, utilizing various paralinguistic feature sets, as follows:

- F0-related features[44,45]: Statistical measures derived from F0 trajectory, jitter, and shimmer.
- Loudness[46]: Statistical features based on the ITU-R BS.1770-4 standard .
- Timbral[47]: Statistical features of hardness, depth, brightness, roughness, warmth, sharpness, and boominess.
- IS-Trait[3]: A comprehensive set of paralinguistic features related to speaker traits as defined in the INTERSPEECH Speaker Trait Challenge 2012.

To isolate the impact of acoustic cues from semantic content, we incorporated a Bag-of-Words (BoW) model as a control variable in the prediction model. The evaluation itself was framed as a binary classification task aimed at predicting both the Big Five traits and the final hiring recommendation, with the F1 score serving as the primary metric. Since the LightGBM classifier consistently outperformed the others across our metrics, we present only

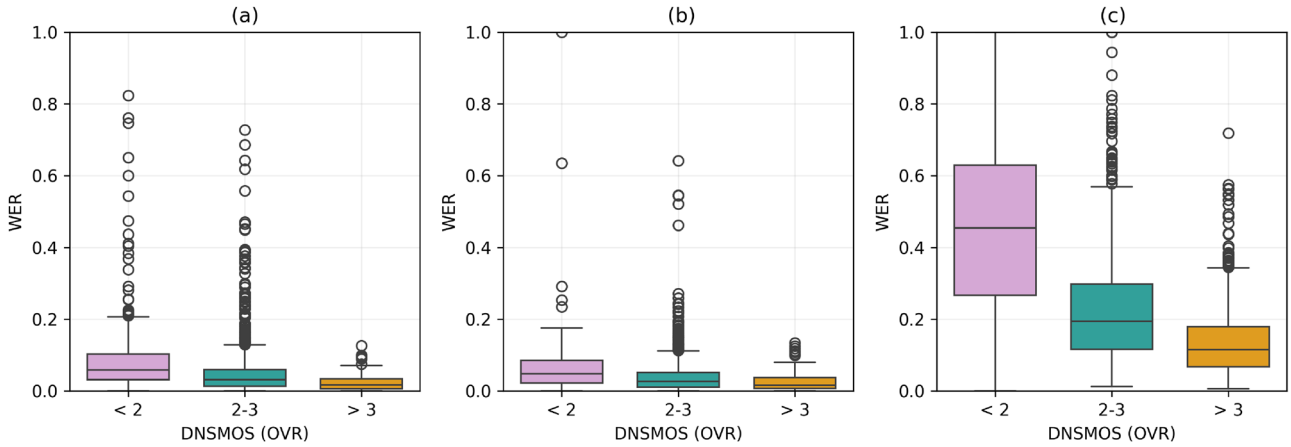

Fig. 2. A plot illustrating the WER, capped at 1.0, for each anonymization method. The data is categorized by DNSMOS (OVR) bins to show the relationship between speech recognition accuracy and estimated audio quality.

Anon. method	F0-related features[44,45]				Loudness[46]				Timbral[47]				IS12-Trait[3]			
	Ex	Co	Ag	Hr	Ex	Co	Ag	Hr	Ex	Co	Ag	Hr	Ex	Co	Ag	Hr
Original	0.794	0.836	0.820	0.655	0.800	0.844	0.810	0.638	0.795	0.850	0.819	0.665	0.783	0.818	0.791	0.681
MAC[29]	0.794	0.832	0.811	0.636	0.804	0.838	0.819	0.634	0.803	0.835	0.816	0.661	0.784	0.823	0.801	0.663
PVT[30]	0.805	0.836	0.821	0.642	0.803	0.842	0.817	0.637	0.800	0.836	0.822	0.662	0.788	0.808	0.803	0.681
NAC[32]	0.786	0.794	0.783	0.609	0.777	0.790	0.781	0.638	0.776	0.789	0.777	0.629	0.773	0.798	0.780	0.692

Table 5. Comparison of speaker trait estimation performance across prosodic features and anonymization methods. Speaker trait abbreviations: Ex (Extraversion), Co (Conscientiousness), Ag (Agreeableness), Hr (Hiring Recommendation).

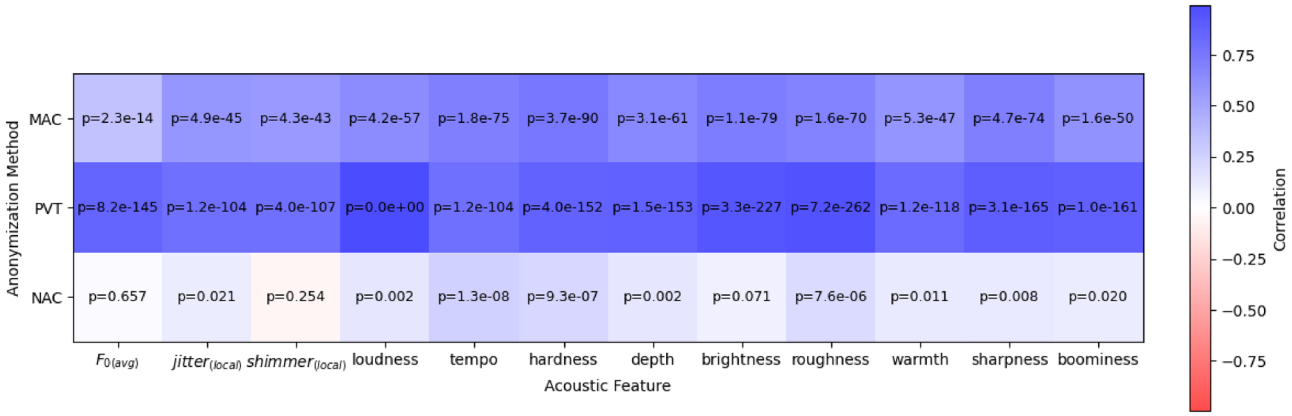

Fig.3. Pairwise correlation analysis of various acoustic features between the original and anonymized speech signals.

its results for clarity and focus. Table 5 provides the representative results for speaker trait estimation, focusing on the top three correlated Big Five traits and the final hiring recommendation.

To understand the trade-off inherent in our methods, we conducted a pairwise correlation analysis between the original and anonymized speech signals to evaluate the preservation of various acoustic features. The results are detailed in Fig. 3. Our analysis reveals that the *PVT* method is highly effective at preserving prosodic information, with a Pearson correlation coefficient greater than 0.5 for most features extracted from the original and anonymized signals. The *MAC* method also shows some level of preservation, as indicated by some darker blue values in the first row of the figure. Conversely, despite offering superior privacy in our ASV evaluation, the *NAC* method demonstrated the lowest preservation of prosodic features. This suggests a trade-off between the privacy level achieved and the fidelity of the original speech's acoustic characteristics.

Our analysis shows no significant performance difference in speaker trait estimation when comparing original speech with speech anonymized by the *MAC* and *PVT* methods, indicating that these approaches successfully preserve utility. In contrast, the *NAC* method generally results in a significant performance drop ($p < 0.005$) for all traits shown in Table 5, with the exception of loudness in the final hiring score prediction. We hypothesize this is because the *NAC* method causes substantial changes to the prosodic information, as illustrated in Fig. 3. As shown in the figure, the correlations for the NAC method are statistically insignificant, with p-values far exceeding the 0.005 threshold across all prosodic features.

The privacy-utility trade-off

The simpler methods, *MAC* and *PVT*, effectively preserve prosodic cues (the how we speak) while minimally altering phonetic content (the what we speak). Crucially, *PVT* maintains sufficient paralinguistic cues to achieve high accuracy in predicting hiring recommendations, demonstrating the viability of computationally efficient methods for this application. Conversely, the *NAC* method achieves superior privacy by significantly disrupting both phonetic details (resulting in high WER) and prosodic information (leading to low trait estimation accuracy and low correlation between original and anonymized paralinguistic features). This confirms that excessive alteration of core acoustic information severely degrades downstream utility.

Our findings underscore the need to preserve key acoustic cues beyond mere linguistic content to maintain utility. Therefore, future work must precisely identify the essential paralinguistic features required for speaker trait perception. This identification will directly guide the advancement of targeted voice anonymization methods. This study evaluated diverse approaches, including classical methods and an early neural codec technique. However, as more advanced neural audio codecs become available, future studies should prioritize evaluating these modern methods. This is necessary to fully understand their evolving privacy-utility trade-offs within the context of robust privacy for personalized Human-Computer Interaction systems.

Conclusion

This study investigated the critical balance between privacy and utility in speaker trait estimation, focusing on the unique challenges of online job interviews. Our research systematically evaluated signal processing-based and neural audio codec-based voice anonymization methods, assessing their ability to remove identifying characteristics while preserving crucial paralinguistic cues for accurate trait prediction.

Our findings reveal that signal processing-based methods, such as the phase vocoder approach, can effectively anonymize speech with only a minor trade-off in utility, maintaining high accuracy in predicting overall hiring recommendations. This demonstrates that a privacy-utility balance is achievable, even in noisy, real-world scenarios. In contrast, while the neural audio codec-based method offers enhanced privacy, it significantly alters some prosodic features, leading to a slight reduction in speaker trait estimation performance. Achieving effective privacy in voice-based systems requires a careful selection of anonymization techniques. The choice depends on the specific application, with simpler signal processing methods being a viable option for many scenarios, while more complex deep learning approaches are necessary for higher privacy needs, albeit with a potential impact on specific utility metrics.

Data availability

The data supporting this study were used under license from third parties[5,13] and are not publicly available. Data may be available upon reasonable request and with third-party permission. Inquiries should be directed to Chee Wee Leong (cleong@ets.org) for the Job Interview Dataset and the Voice Privacy Challenge Organizers (organisers@lists.voiceprivacychallenge.org) for the VPC Dataset.

Received: 30 September 2025; Accepted: 4 February 2026
Published online: 10 February 2026

References

1. Zafar, Z., Paplu, S. & Berns, K. Automatic assessment of human personality traits: A step towards intelligent human-robot interaction. In *IEEE-RAS 18th international conference on humanoid robots* 1–9, https://doi.org/10.1109/HUMANOIDS.2018.8624975 (2018).
2. Song, S. et al. Self-supervised learning of person-specific facial dynamics for automatic personality recognition. *Trans. Affect. Comput.* **14**, 178–195. https://doi.org/10.1109/TAFFC.2021.3064601 (2021).
3. Schuller, B. W. *et al.* The INTERSPEECH 2012 Speaker Trait Challenge. In *Proc. of INTERSPEECH, Portland, Oregon, USA, September 9-13*, 254–257, https://doi.org/10.21437/INTERSPEECH.2012-86 (ISCA, 2012).
4. Schuller, B. et al. A Survey on perceived speaker traits: Personality, likability, pathology, and the first challenge. *Comput. Speech Lang.* **29**, 100–131. https://doi.org/10.1016/j.csl.2014.08.003 (2015).
5. Chen, L. *et al.* Automated video interview judgment on a large-sized corpus collected online. In *Proc. of ACII, San Antonio, TX, USA, October 23-26*, 504–509, https://doi.org/10.1109/ACII.2017.8273646 (IEEE, 2017).
6. Mehta, Y., Majumder, N., Gelbukh, A. & Cambria, E. Recent trends in deep learning based personality detection. *Artif. Intell. Rev.* **53**, 2313–2339. https://doi.org/10.1007/s10462-019-09770-z (2019).
7. Shen, Z., Elibol, A. & Chong, N. Understanding nonverbal communication cues of human personality traits in human-robot interaction. *IEEE/CAA J. Autom. Sinica* **7**, 1465–1477. https://doi.org/10.1109/JAS.2020.1003201 (2020).
8. Kosinski, M., Stillwell, D. & Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci.* **110**, 5802–5805 (2013).
9. Chittaranjan, G., Blom, J. & Gática-Pérez, D. Mining large-scale smartphone data for personality studies. *Pers. Ubiquit. Comput.* **17**, 433–450. https://doi.org/10.1007/s00779-011-0490-1 (2013).
10. Wiedemann, K. Automated processing of personal data for the evaluation of personality traits: Legal and ethical issues. *Eur. Public Law: EU eJ.* https://doi.org/10.2139/ssrn.3102933 (2018).
11. Bäckström, T. Privacy in speech technology. *CoRR* (2025). arXiv2305.05227.
12. Srivastava, B. M. L. et al. Privacy and utility of x-vector based speaker anonymization. *IEEE ACM Trans. Audio Speech Lang. Process.* **30**, 2383–2395. https://doi.org/10.1109/TASLP.2022.3190741 (2022).
13. Tomashenko, N. A. *et al.* The VoicePrivacy 2024 Challenge Evaluation Plan. *CoRR***abs/2404.02677** (2024).
14. Barrick, M. R., Mount, M. K. & Judge, T. A. Personality and performance at the beginning of the new millennium: What do we know and where do we go next?. *Int. J. Sel. Assess.* **9**, 9–30. https://doi.org/10.1111/1468-2389.00160 (2001).
15. McCrae, R. R. & Costa, P. T. Jr. *Personality in adulthood: A five-factor theory perspective* (Guilford Press, 2003).
16. Kang, W. & Malvaso, A. Associations between personality traits and areas of job satisfaction: Pay, work itself, security, and hours worked. *Behav. Sci. (Basel)* **13**, 445. https://doi.org/10.3390/bs13060445 (2023).
17. Polzehl, T., Möller, S. & Metze, F. Automatically assessing personality from speech. In *The 4th international conference on semantic computing* 134–140, https://doi.org/10.1109/ICSC.2010.41 (2010).
18. Gilpin, L. H., Olson, D. M. & Alrashed, T. Perception of speaker personality traits using speech signals. In *Extended Abstracts of the CHI Conference, Montreal, QC, Canada, April 21-26*, https://doi.org/10.1145/3170427.3188557 (ACM, 2018).
19. Pathak, M. A. *Privacy-preserving machine learning for speech processing* (Springer Science & Business Media, 2012).
20. Yoo, I. et al. Speaker anonymization for personal information protection using voice conversion techniques. *IEEE Access* **8**, 198637–198645. https://doi.org/10.1109/ACCESS.2020.3035416 (2020).
21. Panariello, M. et al. The VoicePrivacy 2022 challenge: Progress and perspectives in voice anonymisation. *IEEE ACM Trans. Audio Speech Lang. Process.* **32**, 3477–3491. https://doi.org/10.1109/TASLP.2024.3430530 (2024).
22. Teixeira, F., Abad, A., Raj, B. & Trancoso, I. Privacy-oriented manipulation of speaker representations. *IEEE Access* **12**, 82949–82971. https://doi.org/10.1109/ACCESS.2024.3409067 (2024).
23. Pelikan, M. *et al.* Federated learning with differential privacy for end-to-end speech recognition. *CoRR***abs/2310.00098**, https://doi.org/10.48550/ARXIV.2310.00098 (2023).
24. Arasteh, S. T. *et al.* Differential privacy for protecting patient data in speech disorder detection using deep learning. *ArXiv***abs/2409.19078**, https://doi.org/10.48550/arXiv.2409.19078 (2024).
25. Tayebi Arasteh, S. et al. Addressing challenges in speaker anonymization to maintain utility while ensuring privacy of pathological speech. *Commun. Med.* **4**, 182. https://doi.org/10.1038/s43856-024-00609-5 (2024).
26. Teixeira, F., Abad, A., Raj, B. & Trancoso, I. Privacy-preserving Machine Learning for Remote Speech Processing. In *Proc. IberSPEECH*, 246–250, https://doi.org/10.21437/IberSPEECH.2024-50 (2024).
27. Vecino, B. T. *et al.* Universal Semantic Disentangled Privacy-preserving Speech Representation Learning. *CoRR* https://doi.org/10.48550/ARXIV.2505.13085 (2025). arXiv2505.13085.

28. Wang, S., Chen, Z., Lee, K. A., Qian, Y. & Li, H. Overview of speaker modeling and its applications: From the lens of deep speaker representation learning. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* **32**, 4971–4998. https://doi.org/10.1109/TASLP.2024.3492793 (2024).
29. Patino, J., Tomashenko, N. A., Todisco, M., Nautsch, A. & Evans, N. W. D. Speaker Anonymisation Using the McAdams Coefficient. In *Proc. of Interspeech, Brno, Czechia, August 30 - September 3*, 1099–1103, https://doi.org/10.21437/INTERSPEECH.2021-1070 (ISCA, 2021).
30. Mawalim, C. O., Okada, S. & Unoki, M. Speaker anonymization by pitch shifting based on time-scale modification. In *Proc. 2nd Symposium on Security and Privacy in Speech Communication*, 35–42, https://doi.org/10.21437/SPSC.2022-7 (2022).
31. Meyer, S. *et al.* Speaker anonymization with phonetic intermediate representations. *ArXiv***abs/2207.04834**, https://doi.org/10.48550/arXiv.2207.04834 (2022).
32. Panariello, M., Nespoli, F., Todisco, M. & Evans, N. W. D. Speaker anonymization using neural audio codec language models. In *Proc. of ICASSP, Seoul, Republic of Korea, April 14-19*, 4725–4729, https://doi.org/10.1109/ICASSP48485.2024.10447871 (IEEE, 2024).
33. Défossez, A., Copet, J., Synnaeve, G. & Adi, Y. High fidelity neural audio compression. *Trans. Mach. Learn. Res.* (2023).
34. Costa, P., Terracciano, A. & McCrae, R. Gender differences in personality traits across cultures: Robust and surprising findings. *J. Pers. Soc. Psychol.* **81**(2), 322–331. https://doi.org/10.1037/0022-3514.81.2.322 (2001).
35. Vianello, M., Schnabel, K., Sriram, N. & Nosek, B. Gender differences in implicit and explicit personality traits. *Personality Individ. Differ.* **55**, 994–999. https://doi.org/10.2139/SSRN.2249080 (2013).
36. Srivastava, B. M. L. *et al.* Design Choices for X-Vector Based Speaker Anonymization. In *Proc. of Interspeech, Virtual Event, Shanghai, China, October 25-29*, 1713–1717, https://doi.org/10.21437/INTERSPEECH.2020-2692 (ISCA, 2020).
37. Chen, S. *et al.* VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *CoRR* https://doi.org/10.48550/ARXIV.2406.05370 (2024).
38. Suno-AI. Bark: Text-prompted generative audio model. https://github.com/suno-ai/bark (2023).
39. Desplanques, B., Thienpondt, J. & Demuynck, K. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Proc. of Interspeech*, 3830–3834 (ISCA, 2020).
40. Poddar, A., Sahidullah, M. & Saha, G. Quality measures for speaker verification with short utterances. *Digit. Signal Process.* **88**, 66–79. https://doi.org/10.1016/j.dsp.2019.01.023 (2019).
41. Reddy, C. K. A., Gopal, V. & Cutler, R. DNSMOS P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors. In *Proc. of ICASSP, Virtual and Singapore, 23-27 May*, 886–890, https://doi.org/10.1109/ICASSP43922.2022.9746108 (IEEE, 2022).
42. Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In *Proc. of ICASSP, South Brisbane, Australia, April 19-24*, 5206–5210, https://doi.org/10.1109/ICASSP.2015.7178964 (IEEE, 2015).
43. Mishra, R., Barnwal, S., Malviya, S., Mishra, P. & Tiwary, U. S. *Prosodic Feature Selection of Personality Traits for Job Interview Performance*, 673–682 (Springer, 2020).
44. Cheveigné, A. & Kawahara, H. YIN, A fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* **111**, 1917–30. https://doi.org/10.1121/1.1458024 (2002).
45. Farrús, M., Hernando, J. & Ejarque, P. Jitter and shimmer measurements for speaker recognition. In *Proc. of INTERSPEECH, Antwerp, Belgium, August 27-31*, 778–781, https://doi.org/10.21437/INTERSPEECH.2007-147 (ISCA, 2007).
46. Steinmetz, C. J. & Reiss, J. D. pyloudnorm: A simple yet flexible loudness meter in Python. In *AES Convention* (2021).
47. Pearce, A. *et al.* Second prototype of timbral characterisation tool for semantically annotating non-musical content. Deliverable Report D5.6, AudioCommons (2018).

Author contributions
C.M. and S.O. conceived the experiments, C.M. conducted the experiments, C.M. and C.L. analysed the results. All authors reviewed the manuscript.

Funding
This work was partially supported by JSPS KAKENHI (No. 25K21245, 22H00536, 23H03506), JST Moonshot R&D program (JPMJMS2031), and JST CRONOS (JPMJCS24K7).

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to C.O.M.