Towards Privacy-Preserving Fine-Tuning: Anonymization of Aphasic Speech for Effective ASR

Sebastian Hofstetter, Timo Baumann
Faculty of Computer Science and Mathematics, OTH Regensburg, Germany
sebastian.hofstetter@st.oth-regensburg.de, timo.baumann@oth-regensburg.de

Abstract

The scarcity of publicly available aphasic speech data, driven largely by privacy concerns, poses a significant barrier for fine-tuning Automatic Speech Recognition (ASR) systems in this domain. This study investigates the privacy–utility trade-off of speech anonymization as a strategy to increase data availability. A signal-based McAdams anonymization method is applied to a subset of the AphasiaBank corpus comprising approximately 132 hours of speech from 425 individuals. Privacy is evaluated using an ECAPA-TDNN based Automatic Speaker Verification system and the Equal Error Rate metric. Linguistic utility is assessed by the Word Error Rate using wav2vec2.0 ASR model, tested in multiple conditions, both pretrained and fine-tuned on unprotected and anonymized audio. Our results show that fine-tuning on anonymized aphasic speech data improves ASR performance by 18 % compared to the performance of generic models on non-anonymized speech. Crucially, this gain in utility is achieved alongside substantial privacy protection, with anonymization increasing the privacy by +440 % compared to sharing unprotected speech. This work thus provides a proof-of-concept, demonstrating that speech anonymization mitigates privacy risks to tackle data scarcity and support the development of more effective ASR systems for people with aphasia.

Keywords: Automatic Speech Recognition (ASR), Speech Anonymization, Aphasia

1. Introduction

Automatic Speech Recognition (ASR) has seen major improvements in recent years. However, a notable gap remains in its application to pathological speech, for example to aphasic speech by people with aphasia (PWA). Aphasia is an acquired language disorder resulting from brain injury, most commonly a stroke (80 % of aphasia patients, Rykova and Walther, 2024), which can affect various language system components, such as phonology, morphology, semantics, syntax, and pragmatics (Kohlschein et al., 2017; Sheppard and Sebastian, 2021) while typically sparing motor abilities for reading, writing, or listening (Kohlschein et al., 2017). Roughly one-third of all stroke patients develop aphasia (Sheppard and Sebastian, 2021), with about 20 % suffering from persistent symptoms (Rykova and Walther, 2024). Stroke incidence has risen substantially in the last decades with 11.9 million new stroke cases and 93.8 million people living with the effects of a stroke in the year 2021 alone (Feigin et al., 2025). This illustrates the relevance of this demographic group, as a growing number of strokes leads to more people with aphasia.

Current aphasia diagnosis and therapy are time-consuming and resource-intensive, often hindered by limited access to qualified personnel, especially during the critical post-stroke recovery phase (Wang et al., 2024; Rykova and Walther, 2024). Effective rehabilitation for PWA depends on accurate diagnosis and targeted Speech-Language Therapy (SLT), the intensity and duration of which are key predictors of recovery success (Rykova and Walther, 2024; Kohlschein et al., 2017). Yet, many patients receive insufficient care, which jeopardizes the success of rehabilitation and thus the quality of life of those affected (Wang et al., 2024). Moreover, manual assessment of an individual's aphasia condition is prone to subjectivity, as diagnoses

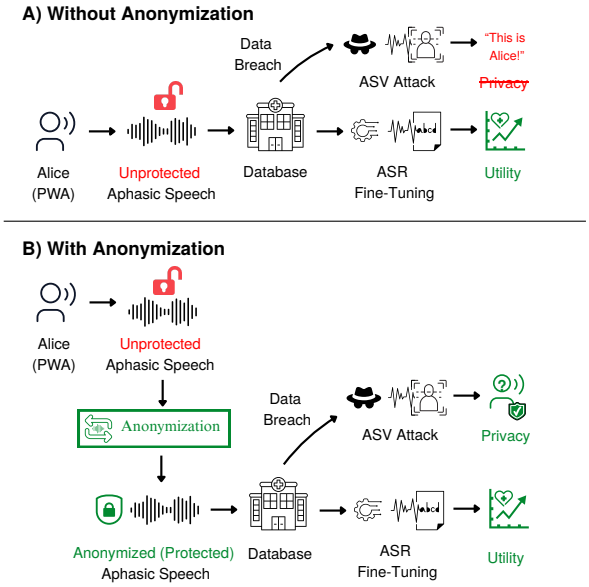

Figure 1: The study's core concept: speech anonymization preserves privacy while maintaining utility when fine-tuning an ASR system. Inspired by Tayebi Arasteh et al. (2024).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5666–5676
11–16 May 2026.

may differ depending on the clinicians' experience (Kohlschein et al., 2017). Therefore the PWA demographic could benefit greatly from advanced technological support and digital applications. Particularly automated systems offer a promising solution to supplement traditional in-person SLT and improve accessibility and diagnostic quality (Rykova and Walther, 2024). A key technology for this purpose is ASR, an application area also referred to as Automatic Disordered Speech Recognition (ADSR) in the context of speech and language disorders (Gohider and Basir, 2024). However, a major challenge is the limitation of state-of-the-art ASR systems being trained exclusively on typical speech. This creates a domain shift problem, as they fail to generalize to the atypical and heterogeneous speech patterns of pathological speech if such data is not present during training (Tobin et al., 2025; Gohider and Basir, 2024). This is especially true for speech from PWA, which shows great diversity, with symptoms and impairments differing widely between syndromes – from global impairments across all language domains to selective difficulties in word retrieval (Wang et al., 2024; Gohider and Basir, 2024). However, research confirms that even small amounts of atypical speech in the training set of ASR models can substantially improve model accuracy on disordered speech for off-the-shelf ASR systems (Tobin et al., 2025). While PWA could benefit significantly from reliable ADSR systems, the models currently available are poorly equipped to meet their specific needs. This study addresses the central obstacle towards the development of such systems: the scarcity of publicly available aphasic speech data.

2. Background and Related Work

Data scarcity remains one of the most persistent challenges in ADSR and particularly aphasia (Wang et al., 2024; Gohider and Basir, 2024; Tayebi Arasteh, 2024). It encompasses both the limited quantity and the lack of diversity in available datasets, which fail to capture the variability of disordered speech (Gohider and Basir, 2024). This issue is particularly present in low-resource languages, where mitigation strategies such as cross-lingual transfer learning (Chatzoudis et al., 2022) do not increase the availability of pathological speech data and have proven to be unreliable in some cases (Mühlhausen et al., 2025).

The primary reason for this scarcity is linked to the strict data protection landscape. Regulations like the GDPR restrict the sharing of pathological speech (Gohider and Basir, 2024; Tayebi Arasteh et al., 2024), which is considered biometric data due to its potential to identify individuals, e. g. through Automatic Speaker Verification (ASV) systems (Nautsch et al., 2019). Pathological speech is even more vulnerable to reveal the speaker identity due to its distinctiveness and thus requires enhanced protection (Tayebi Arasteh et al., 2023). Consequently, large and diverse corpora essential for robust ASR fine-tuning remain scarce.

Researchers have explored privacy-preserving strategies such as speech anonymization to address these challenges (Tayebi Arasteh, 2024). The Voice Privacy (VP) initiative and the associated Voice Privacy Challenge (VPC) format (Tomashenko et al., 2020) established standardized frameworks and benchmarks for evaluating speech anonymization methods. Among these, the McAdams-based anonymization (Patino et al., 2021) has shown particularly promising privacy–utility results when applied to disordered speech (Tayebi Arasteh, 2024). It manipulates speaker-specific spectral features by shifting formant positions derived through Linear predictive coding (LPC) and transformed with a McAdams coefficient (McAdams, 1984). This controlled modification aims to effectively conceal speaker identity while maintaining intelligibility. Recent work by Tayebi Arasteh et al. (2024) demonstrated that McAdams anonymization can successfully balance privacy and utility for several speech disorders, such as Dysarthria, Dysglossia, Dysphonia and Left-Clip and Palate, though results varied by disorder. They called for further research on other speech and language disorders, motivating the present study's focus on aphasia. While the content that is spoken may also contain sensitive information, this aspect is not addressed in our study.

We follow the VPC framework and evaluate utility and privacy using established metrics. Utility is measured by WER, a standard ASR metric representing the ratio of transcription errors to reference words with lower WER scores indicating better performance. Accordingly, the study's primary utility interest is Speech-to-Text (STT) performance improvement through ASR fine-tuning (Tomashenko et al., 2024), while other studies approach utility from different perspectives, such as preserving pathological features for disorder classification Tayebi Arasteh et al. (2024).

We chose wav2vec2.0 (Baevski et al., 2020) as ASR model as it has proven effective in prior studies on similar datasets (Torre et al., 2021). It is based on a self-supervised architecture that processes raw waveforms through a convolutional encoder and a tokenizer that leverages a transformer-based context network and a linear projection layer to learn transcription mappings, fine-tuned using the CTC loss (Baevski et al., 2020; Graves et al., 2006; von Platen, 2021).

Privacy is evaluated using a text-independent ASV system based on the ECAPA-TDNN architec-

ture (Desplanques et al., 2020). The ASV model extracts speaker embeddings of trial utterances and compares them against enrolled speaker profiles using cosine similarity to predict whether the trial samples originate from the same speaker (Mobiny and Najarian, 2018; Reynolds and Rose, 1995). A similarity threshold is necessary to draw a verification decision and the accuracy of this process is expressed through the Equal Error Rate (EER). It quantifies the anonymization strength by identifying the point where false acceptances (trial utterances incorrectly classified as same speaker) and rejections (trial utterances incorrectly classified as different speaker) are equal and thus a stable threshold for the speaker class prediction is reached (Hansen and Hasan, 2015). Higher EER values hinder identification and hence indicate better anonymization. The VPC defines several speaker verification attack conditions to test anonymization robustness. The attack levels unprotected (OO), ignorant (OA) and lazy-informed (AA) are taken into account, representing increasing levels of adversarial knowledge about the anonymization process (Tomashenko et al., 2020). The ignorant attacker has no knowledge about the anonymization system and uses a pre-trained ASV model to create speaker embeddings on unprotected data to perform speaker verification on anonymized data. The lazy-informed attacker has knowledge about which anonymization technique is used by the defender and leverages this information to anonymize the enrollment utterances before creating speaker embeddings to reduce the mismatch caused by anonymization for performing speaker verification on the anonymized data. Yet, still a pre-trained ASV model is employed for the lazy-informed attack level. The unprotected condition serves as a baseline, where no anonymization is applied and the pre-trained ASV model is used to perform speaker verification on unprotected data (Tomashenko et al., 2020).

Despite the recognized challenge of data scarcity in pathological speech analysis and the successful application of anonymization to other speech disorders, there remains a notable gap in research specifically addressing aphasic speech. The unique and heterogeneous nature of aphasic speech, with its diverse subtypes and varying degrees of severity, requires a dedicated investigation into the effectiveness of privacy-preserving techniques. Therefore, the central research question is: *How can aphasic speech data be anonymized for fine-tuning ASR models effectively without compromising patient privacy?* To answer this question, a proof-of-concept study is conducted to demonstrate the successful application of McAdams anonymization on a large, heterogeneous corpus of aphasic speech data from the AphasiaBank database (MacWhinney et al., 2011). The research is guided by the following hypotheses:

Privacy Hypothesis: The anonymization of aphasic speech data will significantly increase the privacy level, as measured by a higher EER for a speaker verification system, compared to non-anonymized data.

Utility Hypothesis: Fine-tuning an ASR model with anonymized aphasic speech will significantly improve utility, as measured by a lower WER, compared to the performance of a pre-trained model.

Privacy-Utility Trade-off Hypothesis: The strength of the anonymization, as adjusted by the McAdams coefficient, will have a negative correlation with the ASR performance, with stronger anonymization yielding lower improvements in WER after fine-tuning.

This work contributes to the research field by providing a succesfull framework for privacy-preserving fine-tuning of ADSR models to encourage the research community to apply anonymization techniques to aphasic speech, thereby helping to overcome the limitations imposed by privacy concerns and data scarcity. This ultimately aims to support the development of more effective ASR systems for PWA. The study's core concept is illustrated in Figure 1.

3. Methodology

3.1. Data Acquisition and Description

The primary dataset for this research is drawn from the AphasiaBank database [1] (MacWhinney et al., 2011). AphasiaBank is a comprehensive repository of multimodal, multilingual materials, including video and audio recordings and transcripts in the Codes for the Human Analysis of Transcripts (CHAT) format (MacWhinney, 2000), from people with aphasia.

From the AphasiaBank English Protocol data, a subset of 909 recordings from 425 distinct speakers with aphasia was selected. The demographic profile of the selected speakers comprises 259 males and 166 females (approximately a 60 %/40 % split), with a mean (median) age of 61 (63). The age range spans 66 years, from 25 to 91 years old. Only speech from PWA was used to ensure the focus remained on pathological speech, excluding those from interviewers or third parties. This resulted in approximately 132 h of pure aphasic speech audio for experimental use. For the purpose of aligning audio and ground truth transcripts, an utterance is defined as a single row in the corresponding CHAT transcript, each annotated with start and end timestamps for efficient audio slicing. The audio recordings were sampled at 16 kHz and stored in WAV format.

[1] https://talkbank.org/aphasia/

The dataset is considered representative of a diverse pathological range of PWA, fulfilling the requirements of a heterogeneous corpus (Wang et al., 2024; Gohider and Basir, 2024). It includes speakers with both non-specific diagnoses (n=65) and specific diagnoses (n=360) according to the Western Aphasia Battery (WAB), encompassing various aphasia types: Severe (Global, n=6), Moderately severe (Broca's, n=115), Moderate (Conduction, n=67; Wernicke's, n=28; Transcortical motor, n=13; Transcortical sensory, n=1), and Mild (Anomic, n=130) according to Kang et al. (2010). The severity of language impairment is further quantified by the WAB Aphasia Quotient (AQ). The dataset reports AQ scores for 395 patients, ranging from a minimum of 10.8 to a maximum of 98.3, spanning an 87.5 AQ point range on the 0-100 scale (lower values indicate higher impairment), with a mean (median) of 66.9 (70.6). This means that the majority of the data is moderate and mild aphasic speech. A histogram illustrating the distribution of AQ scores is presented in Figure 2. Note that the classification into severity types based on the AQ scores is adapted from Barfod et al. (2013), which slightly differs from the classification based on aphasia types according to Kang et al. (2010).

3.2. Data Preprocessing

The handling of silence and pauses is a critical design choice for both ASV and ASR. The study's preprocessing steps aim to retain natural inter-utterance pauses for their pathological and speaker-discriminative value (Angelopoulou et al., 2018; Ossewaarde et al., 2025). However, extended inter-utterance non-speech segments – e. g., during picture description tasks following the AphasiaBank protocol (MacWhinney et al., 2011) – were excluded to avoid fine-tuning on task-specific artifacts rather than natural aphasic speech patterns, and to prevent degrading ASR model performance. Outliers were identified via Voice Activity Detection (VAD)-based silence filtering using the IQR method (Vinutha et al., 2018) ($1.5 \times$ IQR upper boundary), removing 8.7 % of audio chunks. Segments with overlapping speech were discarded. Length thresholds of $\geq 1.0, s$ (ASR only) and $\geq 1.8, s$ (ASR and ASV) were enforced (Tayebi Arasteh, 2024).

Audio chunks lacking corresponding transcript ground truth were excluded from the ASR pipeline to enable seamless calculation of the WER metric. To ensure a fair WER calculation, the CHAT formatted transcripts were transformed into a format suitable for ASR output comparison. The implemented rules aim to obtain a transcription that is as verbatim as possible. This involved removing CHAT-specific formatting and annotations, such as pause indicators, length annotations, punctuation, and descriptive notes, e. g. motion descriptions.

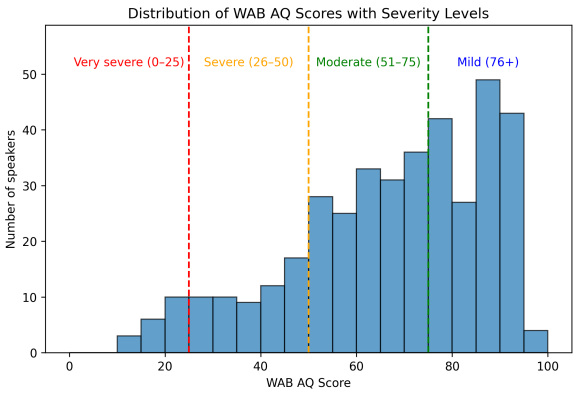

Figure 2: Distribution of WAB AQ scores in our dataset. Scores are grouped in five-point-bins and classified in four severity types (Barfod et al., 2013).

Phonetically transcribed words were replaced with their intended standard spelling (where annotated). A dictionary of common abbreviations and slang words, extracted from the official CHAT manual (MacWhinney, 2000), was utilized to ensure consistent spelling (e.g., "hasta" to "has to" or "wanna" to "want to"). This standardization was chosen after manual inspection revealed that human transcribers often used abbreviations rather than the common spelling. An example of the transcript cleaning is illustrated in Table 1.

3.3. Experiment Design

The evaluation framework adheres to the established guidelines of the Voice Privacy Challenge initiative (Tomashenko et al., 2024). The Equal Error Rate (EER) serves as the primary privacy metric, assessing the effectiveness of anonymization techniques against an ASV system. For ASR utility, we employ the widely accepted WER. These

Table 1: Example of transcript cleaning. The original transcript is in CHAT format, the processed transcript is cleaned and ready for ASR evaluation.

Original (CHAT-formatted) Transcript:	&-uh oh_my_god ye:s, so &=chuckles I moved to hæləf rnjə@u [: California] [* p:n] with my &+mo [/] &-uhm my [//] mother [: daughter] (.) &=points:daughter, so I θe @u [: x@n] coulda &+re recovery (..) (okay)
Processed Transcript:	uh oh my god yes so I moved to California with my mo uhm my mother so I could have re recovery okay

metrics strengthen the comparability of our study. This study focuses solely on linguistic content for utility evaluation, excluding emotional content.

Inspired by the findings of Tayebi Arasteh (2024), representing state-of-the-art in anonymization for pathological speech and demonstrating superior performance of signal-based over deep learning-based techniques, this study utilizes the McAdams anonymization method., this study utilizes the McAdams anonymization method. Given the resource constraints of this study, only this technique is employed to demonstrate a proof of concept for effective anonymization of aphasic speech, laying the foundation for future optimization of the privacy–utility trade-off. The VPC 2022 baseline (B2) McAdams anonymization implementation,[2] is used (Tomashenko et al., 2022a), with McAdams coefficients randomly sampled between 0.5 and 0.9 on a per-utterance basis, consistent with Tomashenko et al. (2024).

To ensure fair and realistic ASR evaluation, the "hold-speakers-out" principle is strictly followed for splitting data into training, validation, and test sets with a 70/15/15 split ratio. This approach guarantees that each speaker contributes to only one set, promoting higher generalization to unseen speakers (Liu et al., 2023). The study aims for speaker-independent ASR, enhancing model robustness against speaker-specific traits and allowing more extensive application compared to personalized models (Tobin et al., 2025). To mitigate the risk of training on homogenous sets, which could lead to poor performance on diverse aphasia types, the experiment leverages a large sample size of over 400 unique speakers across all sets, with 260 speakers in the training set alone. In addition, a nuanced partitioning strategy was employed, grouping patients by aphasia type, severity, gender, and age to balance these characteristics across sets, followed by random sampling within each group.

For ASR fine-tuning and evaluation, a boundary silence padding was applied to prevent abrupt cutoffs from sharp timestamp-based audio slicing, following the rationale of Gulli et al. (2024). VAD was used to measure existing silence at utterance boundaries, trimming or extending boundary silence as needed to standardize padding to 250 ms.

The fine-tuning implementation is inspired by the guidelines outlined by von Platen (2021). Two fine-tuned models were developed, one with original audio and the second with anonymized audio, starting from a pretrained checkpoint provided by Meta on Hugging Face [3]. Hyperparameter optimization (learning rate, batch size, epochs) was performed using a grid search on 1 % of the training set and measuring the WER on the validation set, yielding optimal settings of a learning rate of 2×10^{-5}, a batch size of 2, and 50 epochs for both models. For the original dataset fine-tuning, low regularization was applied, consistent with Torre et al.'s fine-tuning hyperparameters on a similar dataset (Torre et al., 2021). Feature and layer dropout rates were fixed at 0.05 and 0.02, while activation and attention dropout were set lower at 0.03 and 0.036. The temporal mask was set to 0.057, and gradient accumulation was done after each step. Fine-tuning on anonymized data proved less stable, leading to a simplified setup with a uniform dropout rate of 0.01 by manual search. Both models employed a weight decay of 0.005, 1000 warmup steps, and a linear decay learning rate scheduler.

Following the VPC framework, the ASV setup simulates an attacker with access to enrollment utterances. ASV baseline data homogeneity can increase the EER (Tayebi Arasteh et al., 2023). Therefore, in order to establish the EER as a lower privacy bound and avoid narrowing the privacy statement to specific subgroups, this study opts for a random selection of ASV speaker pairs, disregarding demographic or pathological traits like gender, age or aphasia type.

In addition, bootstrap sampling was employed to generate 50 unique evaluation runs per attack level. In each run, speakers were randomly drawn from the entire set, allowing for variation in the weighting of aphasia types and strengths, thus simulating attacks on diverse databases. The final EER is reported as the mean of these bootstrap runs, with the standard deviation indicating the spread. The emplyoed ECAPA-TDNN ASV model was pre-trained on Voxceleb1 and Voxceleb2 training datasets and provided by Speechbrain on HuggingFace [4] (Ravanelli et al., 2021) and was not fine-tuned on aphasic speech or McAdams-anonymized data.

For each speaker, enrollment and trial utterances were defined, with a same-speaker ratio of approximately 5 % in the trial set (Tomashenko et al., 2024). This aims to mirror the imbalance in larger datasets, where attackers likely encounter far more recordings from different speakers than from the target speaker. The enrollment set consists of at least five utterances and no more than 15 % of total utterances per speaker (Meyer et al., 2024) with a minimum of $1.8\,s$ duration per utterance (Tayebi Arasteh, 2024). Although it cannot be guaranteed that all trial utterances stem from recordings unknown to the attacker for all speakers, given that 260 speakers only have one recording, this principle is applied for 165 speakers with more than one recording to

[2] https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2022

[3] https://huggingface.co/facebook/wav2vec2-base-960h

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

minimize potential same-conversation bias. Furthermore, all enrollment utterances were excluded from trial utterance candidates.

To examine the relationship between anonymization strength and ASR performance degradation, WER differences were calculated as $WER\_anonymized - WER\_original$ for each audio sample in the test set. This approach aims to isolate the effect of anonymization and minimize influence by sample specific traits, e. g. the speech inteligibility due to aphasia severity, aphasia type or recording quality. The correlation is assessed for fine-tuned model results only, assuming higher relevance in real world application. The McAdams coefficient values were binned into 0.05-unit intervals, and mean WER differences (i.e., increases) were computed for each bin. The interpretation primarily focused on three representative anonymization levels: McAdams values of 0.5 (strong anonymization), 0.7 (moderate anonymization), and 0.9 (weak anonymization).

3.4. Statistical Analysis

For the privacy hypothesis, Shapiro-Wilk tests were conducted for all attack levels to assess the normality of EER values. As the values were found to be normally distributed, two one-sided paired t-tests were performed to determine the significance of differences for the ignorant and lazy-informed attack levels compared to the unprotected baseline (Field et al., 2012).

For the utility hypothesis, visual assessment using histograms and Q-Q plots revealed non-normal distributions for the individual WER values. Therefore, Wilcoxon signed-rank tests were employed to assess statistical significance between the means (Field et al., 2012).

The privacy-utility trade-off was assessed on the individual utterance basis. The McAdams coefficient served as the measure of privacy strength, as lower coefficients correlate with higher EER values due to greater signal manipulation (Patino et al., 2021). WER was used as the measure of utility.

To ensure the appropriate statistical methods were used, the data distributions were examined. The Kolmogorov-Smirnov test (Walther, 2022) indicated significant departures from normality for both McAdams anonymization values ($D = 0.0574, p < .001$) and WER differences ($D = 0.185, p < .001$) in the test set (N = 14,103). Visual assessment through histograms and Q-Q plots confirmed the statistical tests, showing clear deviations from normal distribution patterns. Given these violations of parametric assumptions, Spearman's rank correlation was employed to assess the monotonic relationship between variables, as it is robust to non-normal distributions (Field et al., 2012). A significance level of $p \leq 0.05$ was used for all analyses.

3.5. Reproducibility

The code repository, containing the implementation of the experiment and information about the hardware and software used, is publicly available [5].

4. Results and Discussion

4.1. Privacy Gains of Anonymization

For privacy evaluation, we compared the EER of the baseline attack level (original audio conditions, OO) against the attack levels involving anonymized audio conditions (OA and AA). Table 2 summarizes the average EER values across all trials and the respective privacy gains of +440 % and +640 %, showing substantial increases for both attack levels against the baseline, thus supporting the privacy hypothesis. Paired *t*-tests confirmed that the anonymization approach significantly increased the EER compared to the unprotected baseline. The ignorant attack condition (OA) resulted in a mean increase of $+0.181$ ($t(49) = 215.9$, $p < .001$), while the lazy-informed attack condition (AA) showed a mean increase of $+0.125$ ($t(49) = 148.6$, $p < .001$). These results indicate that the McAdams anonymization method provides a statistically significant privacy gain for both attack levels, with the ignorant attacker yielding a stronger effect than the lazy-informed attacker. An Equal Error Rate of 2.9 % for the baseline unprotected data attack level indicates that aphasic speech is highly vulnerable to speaker verification attacks. This is a meaningful finding when compared to non-pathological speech from datasets like LibriSpeech (Panayotov et al., 2015), which typically show higher EER values and thus less vulnerability to speaker verification attacks (Tomashenko et al., 2024). At the same time, when compared to the findings of Tayebi Arasteh et al.

[5] https://github.com/sobs0/aphasia-anonym

Table 2: EER comparison for different attack levels (unprotected, ignorant, lazy-informed; Tomashenko et al., 2020), including the privacy gain compared to the unprotected baseline.

Attack Level	Pretrained ASV Model	Privacy Gain
Unprotected (OO)	2.83 $\pm$ 0.24	–
Ignorant (OA)	21.0 $\pm$ 0.58	+640 %
Lazy-informed (AA)	15.3 $\pm$ 0.55	+440 %

(2024), aphasic speech appears to be less vulnerable than disorders like Dysglossia, Dysarthria, or Dysphonia, but more vulnerable than Left Clip and Palate. However, this comparison is limited due to differences in the ASV model setups.

A further comparison of anonymization results across studies reveals an interesting insight. This study and the VPC used the same ECAPA-TDNN architecture and implementation for ASV (Desplanques et al., 2020; Ravanelli et al., 2021; Tomashenko et al., 2024) and McAdams anonymization implementation (Tomashenko et al., 2022a). This study's EER values for anonymized data (20.96 % for ignorant and 15.30 % for lazy-informed) are partially lower and closer together but relatively similar to those reported in the VPC 2020 (26.17 % for ignorant and 13.14 % for lazy-informed; Tomashenko et al., 2022b). This might be due to the inherent vulnerability of pathological speech that remains even after anonymization. On the other hand, the EER values from the work of Tayebi Arasteh (2024), which used a GE2E architecture (Wan et al., 2018) for ASV, were at least 10 percentage points higher than our weakest attack level. This substantial difference suggests that the ECAPA-TDNN model may be less vulnerable to signal-based anonymization techniques like McAdams, providing a more conservative and potentially safer privacy estimate for aphasic speech.

Furthermore the results show that, although the McAdams anonymization significantly improves the privacy level, the effect decreases with a more knowledgeable attacker, which is consistent with the findings of Tomashenko et al. (2022b).

4.2. Utility of Fine-Tuning

Utility was measured by Word Error Rate (WER) for each individual item of the ASR test set, which consisted of 81 unique speakers and 14103 samples. As shown in Table 3, fine-tuning substantially improved the model's performance on aphasic speech by at least +33 %. The fine-tuned model trained on anonymized data outperformed the pre-trained model on original data by +18.03 %, demonstrating that it is possible to achieve improved utility even with privacy-preserving data. This finding supports the Utility Hypothesis.

A one-sided, paired Wilcoxon signed-rank test (Field et al., 2012) confirmed a significant difference between the WER scores of the fine-tuned model on anonymized data and the pretrained model on original data ($V = 40695124, z = 0.415, p < .001$), with a moderate effect size ($r = 0.41$) (Cohen, 1988). As expected, anonymization caused higher WER values, causing a relative utility decrease of -38.86 % for the pre-trained model and -22.36 % for the fine-tuned model.

4.3. Privacy Utility Correlation

Spearman's rank correlation revealed a statistically significant small negative correlation between McAdams anonymization values and WER differences (Spearman's $\rho = -0.1714, p < .001, N = 14,103$). This indicates that stronger anonymization levels (lower McAdams coefficients) are associated with larger increases in WER, therefore supporting the Privacy-Utility Trade-off Hypothesis.

The relationship is visualized in Figure 3, which shows a clear trend of increasing WER as the McAdams coefficient decreases. The degradation was non-linear and showed an accelerating pattern. For instance, moving from moderate to strong anonymization level (from McAdams coefficient 0.7 to 0.5) resulted in a 116 % increase in WER difference, compared to a 66.5 % increase when moving from weak to moderate anonymization (from McAdams coefficient 0.9 to 0.7) and suggests that the performance penalty becomes disproportionately high at higher anonymization levels. This finding is consistent with previous research, which reported substantial audio quality losses of disordered speech for McAdams coefficient values of 0.75 and below (Tayebi Arasteh et al., 2024). ASR performance is negatively affected by audio quality decrease (Chen, 2009), therefore the statistical correlation and this theoretical link suggest a causal relationship: stronger McAdams anonymization causes a higher decrease in ASR performance for aphasic speech.

5. Conclusion

The scarcity of aphasic speech data, largely due to privacy concerns, is a major bottleneck in the development of effective voice applications for people

Table 3: WER comparison for different audio conditions and models, including performance gains and losses. The framed value shows the relative gain of Anonymized Audio + Fine-tuned Model over Original Audio + Pre-trained Model.

	Pre-trained Model	**Fine-tuned Model**	**Fine-tuning Gain**
Original Audio	62.8 $\pm$ 36.6	42.0 $\pm$ 34.7	+33.0 %
Anonymized Audio	87.2 $\pm$ 29.5	51.4 $\pm$ 34.1	+41.0 %
Anonymization Loss	-38.9 %	-22.4 %	+18.0 %

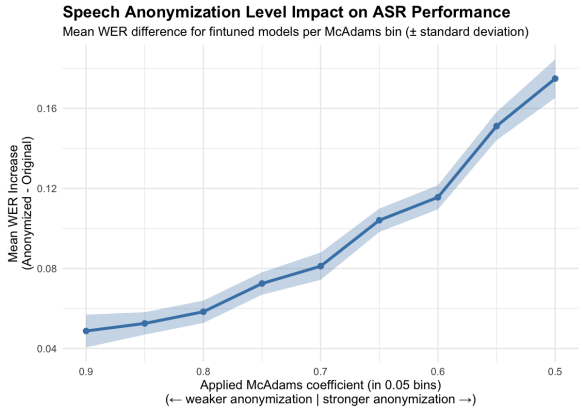

Figure 3: Trendline of McAdams anonymization values and WER increase from evaluation on anonymized to original audio in the fine-tuned ASR model versions. Stronger anonymization yields lower ASR quality.

with aphasia. However, these applications could have a highly positive impact on the diagnosis and treatment of aphasia, presenting the motivation for further investigation on effective ADSR systems. This study addressed this challenge by investigating a privacy-preserving fine-tuning framework using the McAdams anonymization technique on a large, heterogeneous dataset from the AphasiaBank database. Our findings demonstrate a viable path to overcome this limitation by successfully achieving improvements in both privacy and utility.

The results confirm our research hypotheses: **Privacy Hypothesis:** The McAdams anonymization method significantly increased privacy, with EER values rising from a vulnerable baseline of 2.83 % to over 15.3 % for both attack levels, measured using an ECAPA-TDNN ASV model. This demonstrates that a meaningful level of privacy protection is achievable for aphasic speech data. This is of great necessetiy, as the baseline EER value suggests high vulnerability for speaker verification attacks for PWA.

Utility Hypothesis: Fine-tuning a pre-trained wav2vec2.0 model on anonymized aphasic speech significantly improved its performance, with performance gains of over 33.0 % for both fine-tuned versions. This highlights the need and effectiveness of fine-tuning efforts for ADSR systems also in the aphasia domain. The fine-tuned model on anonymized data achieved a WER of 51.5 %. Although that is still a high error level, it outperforms the pre-trained model on original data by 18.0 %, which represents the practice if no aphasic speech data is available for fine-tuning. This finding supports the feasibility of developing more effective ASR systems for aphasia using privacy-preserving data.

Privacy-Utility Trade-off Hypothesis: A significant negative correlation was found between the McAdams coefficient and the increase in WER, indicating that stronger anonymization leads to a greater loss in utility. This highlights the critical need to carefully balance these two factors when designing a privacy-preserving fine-tuning system. The non-linear nature of this trade-off suggests that a minimal increase in privacy comes with a disproportionately large penalty in utility at higher anonymization levels, making it crucial to select a suitable anonymization strength based on the specific application's requirements.

In conclusion, this proof-of-concept study provides a successful framework for the privacy-preserving fine-tuning of ASR models on aphasic speech. The results of this study demonstrate that this approach significantly enhances both privacy and utility. Specifically, a substantial improvement in privacy by at least 440 % was achieved, while simultaneously boosting the ASR utility by 18.03 %.

6. Limitations

This study is limited by the size of the dataset, language, and demographic composition. The data consist solely of English-speaking individuals with aphasia, with a skew towards mild and moderate cases and a slight gender imbalance (60 % male, 40 % female). Future work should validate the findings on larger, multilingual, and more balanced datasets – particularly in low-resource languages – to ensure generalizability and demographic fairness.

Model-wise, the investigation relied on a single ASV model (ECAPA-TDNN) and ASR model (wav2vec2.0). Future studies should confirm these results using alternative architectures, such as GE2E for ASV (Wan et al., 2018) and Whisper for ASR (Radford et al., 2023), to enhance comparability with related work (Tayebi Arasteh, 2024; Mühlhausen et al., 2025). The differing privacy improvements compared to Tayebi Arasteh (2024) also highlight the need to study model-specific vulnerabilities and unify privacy evaluation standards across different experiment setups.

Furthermore, this research focused solely on McAdams anonymization. Future work should explore additional anonymization techniques, particularly irreversible and deep learning-based methods, as the McAdams transformation is deterministic and thus theoretically invertible. While practical recovery is substantially constrained by the per-utterance random sampling of coefficients, this remains an inherent limitation of signal-based approaches that more sophisticated methods may overcome. This also motivates applying more

knowledgeable attack levels, e.g. the semi-informed attack condition (Tomashenko et al., 2024), to strengthen privacy assessment. Given that this study evaluated utility only through WER, alternative definitions of utility should be considered in the future, such as pathological feature preservation for diagnosis. A more detailed analysis of the privacy–utility trade-off on the EER-level would help define optimal anonymization strength for various application requirements. However, it should also be acknowledged that publishing anonymized aphasic speech data comes with the long-term risk of increasingly powerful ASV models potentially compromising privacy guarantees achieved today.

7. References

Georgia Angelopoulou, Dimitrios Kasselimis, George Makrydakis, Maria Varkanitsa, Petros Roussos, Dionysis Goutsos, Ioannis Evdokimidis, and Constantin Potagas. 2018. Silent pauses in aphasia. *Neuropsychologia*, 114:41–49.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 12449–12460, Red Hook, NY, USA. Curran Associates Inc.

Vanessa Barfod, Annabel McDermott, and Nicol Korner-Bitensky. 2013. Western Aphasia Battery (WAB) – Strokengine.

Gerasimos Chatzoudis, Manos Plitsis, Spyridoula Stamouli, Athanasia-Lida Dimou, Nassos Katsamanis, and Vassilis Katsouros. 2022. Zero-Shot Cross-lingual Aphasia Detection using Automatic Speech Recognition. In *Interspeech 2022*, pages 2178–2182. ISCA.

Lei Chen. 2009. Audio quality issue for automatic speech assessment. In *Speech and Language Technology in Education (SLaTE 2009)*, pages 97–100. ISCA.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, second edition edition. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*, pages 3830–3834.

Valery L Feigin, Michael Brainin, Bo Norrving, Sheila O Martins, Jeyaraj Pandian, Patrice Lindsay, Maria F Grupper, and Ilari Rautalin. 2025. World Stroke Organization: Global Stroke Fact Sheet 2025. *International Journal of Stroke*, 20(2):132–144.

Andy Field, Jeremy Miles, and Zoë. Field. 2012. *Discovering Statistics Using R*. Sage, London.

Nada Gohider and Otman A. Basir. 2024. Recent advancements in automatic disordered speech recognition: A survey paper. *Natural Language Processing Journal*, 9:100110.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, pages 369–376, Pittsburgh, Pennsylvania. ACM Press.

Andrea Gulli, Francesco Costantini, Diego Sidraschi, and Emanuela Li Destri. 2024. Fine-Tuning a Pre-Trained Wav2Vec2 Model for Automatic Speech Recognition- Experiments with De Zahrar Sproche. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7336–7342, Torino, Italia. ELRA and ICCL.

John H.L. Hansen and Taufiq Hasan. 2015. Speaker Recognition by Machines and Humans: A tutorial review. *IEEE Signal Processing Magazine*, 32(6):74–99.

Eun Kyoung Kang, Hae Min Sohn, Moon-Ku Han, Won Kim, Tai Ryoon Han, and Nam-Jong Paik. 2010. Severity of Post-stroke Aphasia According to Aphasia Type and Lesion Location in Koreans. *Journal of Korean Medical Science*, 25(1):123–127.

Christian Kohlschein, Maximilian Schmitt, Björn Schüller, Sabina Jeschke, and Cornelius J. Werner. 2017. A machine learning based system for the automatic evaluation of aphasia speech. In *2017 IEEE 19th International Conference on E-Health Networking, Applications and Services (Healthcom)*, pages 1–6.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2023. Investigating data partitioning strategies for crosslinguistic low-resource ASR evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 123–131, Dubrovnik, Croatia. Association for Computational Linguistics.

Brian MacWhinney. 2000. The CHILDES project: Tools for analyzing talk. *Child Language Teaching and Therapy*.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. AphasiaBank: Methods for Studying Discourse. *Aphasiology*, 25(11):1286–1307.

Stephen McAdams. 1984. *Spectral Fusion, Spectral Parsing and the Formation of Auditory Images*. Ph.D. thesis, Stanford University, Stanford, California.

Sarina Meyer, Florian Lux, and Ngoc Thang Vu. 2024. Probing the Feasibility of Multilingual Speaker Anonymization. In *Proc. Interspeech 2024*, pages 4448–4452.

Aryan Mobiny and Mohammad Najarian. 2018. Text-Independent Speaker Verification Using Long Short-Term Memory Networks.

Sara Mühlhausen, Sarah Gomez, Norina Lauer, and Timo Baumann. 2025. Cross lingual transfer learning does not improve aphasic speech recognition. In *Elektronische Sprachsignalverarbeitung 2025: Tagungsband Der 36. Konferenz Halle/Saale, 05.–07. MÄRZ 2025*. TUDpress.

Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, Els Kindt, Héctor Delgado, Massimiliano Todisco, Mohamed Amine Hmani, Aymen Mtibaa, Mohammed Ahmed Abdelraheem, Alberto Abad, Francisco Teixeira, Driss Matrouf, Marta Gomez-Barrero, Dijana Petrovska-Delacrétaz, Gérard Chollet, Nicholas Evans, Thomas Schneider, Jean-François Bonastre, Bhiksha Raj, Isabel Trancoso, and Christoph Busch. 2019. Preserving privacy in speaker and speech characterisation. *Computer Speech & Language*, 58:441–480.

Roelant Ossewaarde, Yolande Pijnenburg, Antoinette Keulen, Roel Jonkers, and Stefan Leijnen. 2025. Role of pause duration in primary progressive aphasia. *Aphasiology*, 39(5):601–619.

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

Jose Patino, Natalia Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas Evans. 2021. Speaker Anonymisation Using the McAdams Coefficient. In *Proc. Interspeech 2021*, pages 1099–1103.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 28492–28518, Honolulu, Hawaii, USA. JMLR.org.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A General-Purpose Speech Toolkit.

D.A. Reynolds and R.C. Rose. 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83.

Eugenia Rykova and Mathias Walther. 2024. AphaDIGITAL – Digital Speech Therapy Solution for Aphasia Patients with Automatic Feedback Provided by a Virtual Assistant. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, pages 3385–3394.

Shannon M. Sheppard and Rajani Sebastian. 2021. Diagnosing and managing post-stroke aphasia. *Expert review of neurotherapeutics*, 21(2):221–234.

Soroosh Tayebi Arasteh. 2024. *Tackling Data Scarcity in Automatic Pathological Speech Analysis*. Doctoralthesis, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Technische Fakultät.

Soroosh Tayebi Arasteh, Tomás Arias-Vergara, Paula Andrea Pérez-Toro, Tobias Weise, Kai Packhäuser, Maria Schuster, Elmar Noeth, Andreas Maier, and Seung Hee Yang. 2024. Addressing challenges in speaker anonymization to maintain utility while ensuring privacy of pathological speech. *Communications Medicine*, 4(1):1–16.

Soroosh Tayebi Arasteh, Tobias Weise, Maria Schuster, Elmar Noeth, Andreas Maier, and Seung Hee Yang. 2023. The effect of speech pathology on automatic speaker verification: A large-scale study. *Scientific Reports*, 13(1):20476.

Jimmy Tobin, Katrin Tomanek, and Subhashini Venugopalan. 2025. Towards a Single ASR Model That Generalizes to Disordered Speech. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

N. Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco. 2020. Introducing the VoicePrivacy Initiative. In *Interspeech 2020*, pages 1693–1697. ISCA.

Natalia Tomashenko, Xiaoxiao Miao, Pierre Champion, Sarina Meyer, Xin Wang, Emmanuel Vincent, Michele Panariello, Nicholas Evans, Junichi Yamagishi, and Massimiliano Todisco. 2024. The VoicePrivacy 2024 Challenge Evaluation Plan.

Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, Hubert Nourtel, Pierre Champion, Massimiliano Todisco, Emmanuel Vincent, Nicholas Evans, Junichi Yamagishi, and Jean-François Bonastre. 2022a. The VoicePrivacy 2022 Challenge Evaluation Plan.

Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, Anaïs Chanclu, Jean-François Bonastre, Massimiliano Todisco, and Mohamed Maouche. 2022b. The VoicePrivacy 2020 Challenge: Results and findings. *Computer Speech & Language*, 74:101362.

Iván G. Torre, Mónica Romero, and Aitor Álvarez. 2021. Improving Aphasic Speech Recognition by Using Novel Semi-Supervised Learning Methods on AphasiaBank for English and Spanish. *Applied Sciences*, 11(19):8872.

H. P. Vinutha, B. Poornima, and B. M. Sagar. 2018. Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset. In *Information and Decision Sciences*, pages 511–518. Springer, Singapore.

Patrick von Platen. 2021. Fine-Tune Wav2Vec2 for English ASR in Hugging Face with Transformers.

Björn Walther. 2022. Kolmogorov-Smirnov-Test in R rechnen.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized End-to-End Loss for Speaker Verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883.

Yin Wang, Weibin Cheng, Fahim Sufi, Qiang Fang, and Seedahmed S. Mahmoud. 2024. A Systematic Review of Using Deep Learning in Aphasia: Challenges and Future Directions. *Computers*, 13(5):117.