


RESEARCH ARTICLE

Obfuscation via pitch-shifting for balancing privacy and diagnostic utility in voice-based cognitive assessment

Meysam Ahangaran¹ | Nauman Dawalatabad^{2,#} | Cody Karjadi^{3,4} | James Glass⁵ | Rhoda Au^{1,3,4,6,7,8} | Vijaya B. Kolachalama^{1,8,9,10} ¹Department of Medicine, Boston University Chobanian and Avedisian School of Medicine, Boston, Massachusetts, USA²Zoom Communications Inc., San Jose, California, USA³Department of Anatomy and Neurobiology, Boston University Chobanian and Avedisian School of Medicine, Boston, Massachusetts, USA⁴The Framingham Heart Study, Boston University Chobanian and Avedisian School of Medicine, Boston, Massachusetts, USA⁵Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA⁶Department of Neurology, Boston University Chobanian and Avedisian School of Medicine, Boston, Massachusetts, USA⁷Department of Epidemiology, Boston University School of Public Health, Boston, Massachusetts, USA⁸Boston University Alzheimer's Disease Research Center, Boston, Massachusetts, USA⁹Department of Computer Science, Boston University, Boston, Massachusetts, USA¹⁰Faculty of Computing and Data Sciences, Boston University, Boston, Massachusetts, USA

Correspondence

Vijaya B. Kolachalama, Department of Medicine, Boston University Chobanian and Avedisian School of Medicine, 72 E. Concord St, Boston, MA – 02118, USA.
Email: vkola@bu.edu

Abstract

INTRODUCTION: Digital voice analysis is an emerging tool for differentiating cognitive states, but it poses privacy risks as automated systems may inadvertently identify speakers.**METHODS:** We developed a computational framework to evaluate the trade-off between voice obfuscation and cognitive assessment accuracy, using pitch-shifting as a representative method. This framework was applied to voice recordings from the Framingham Heart Study (FHS, $n = 128$) and the DementiaBank Delaware (DBD, $n = 85$) corpus, both featuring responses to neuropsychological tests. Speaker obfuscation was measured via equal error rate (EER), and diagnostic utility was assessed through machine learning models distinguishing cognitive states: normal cognition (NC), mild cognitive impairment (MCI), and dementia (DE).**RESULTS:** With the top 20 acoustic features, our framework achieved classification accuracies of 62.2% (EER: 0.3335) on the FHS dataset for NC, MCI, and DE differentiation, and 63.7% (EER: 0.1796) on the DBD dataset for NC and MCI differentiation, using obfuscated speech files.**DISCUSSION:** Our results demonstrate the feasibility of privacy-preserving voice markers, offering a scalable solution for voice-based cognitive assessments.

KEYWORDS

cognitive assessment, digital health, privacy, voice recordings

Highlights

- We developed a computational framework using pitch-shifting and acoustic transformations to balance speaker privacy and diagnostic utility in voice-based cognitive assessments.

Work done while at Massachusetts Institute of Technology.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.© 2025 The Author(s). *Alzheimer's & Dementia* published by Wiley Periodicals LLC on behalf of Alzheimer's Association.

Funding information

National Institute on Aging's Artificial Intelligence and Technology Collaboratories, Grant/Award Numbers: P30-AG073104, P30-AG073105; American Heart Association, Grant/Award Number: 20SFRN35460031; Gates Ventures; National Institutes of Health, Grant/Award Numbers: R01-HL159620, R01-AG062109, R01-AG083735

- We evaluated the framework on two independent datasets, Framingham Heart Study (FHS, $n = 128$) and DementiaBank Delaware (DBD, $n = 85$) corpus, assessing the trade-off between privacy (measured by equal error rate [EER]) and classification accuracy.
- Our framework achieved classification accuracies of 62.2% (EER: 0.3335) for distinguishing normal cognition (NC), mild cognitive impairment (MCI), and dementia in the FHS dataset and 63.7% (EER: 0.1796) for NC and MCI differentiation in the DBD dataset, using obfuscated speech files.
- Our framework demonstrates that pitch-shifting levels can preserve diagnostic utility while protecting speaker identity, offering a scalable and privacy-preserving solution.

1 | BACKGROUND

Digital voice recordings contain valuable information that can indicate an individual's cognitive health,¹ offering a non-invasive and efficient method for assessment. Research has demonstrated that digital voice measures can detect early signs of cognitive decline by analyzing features such as speech rate, articulation, pitch variation, and pauses, which may signal cognitive impairment when deviating from normative patterns.² Advancements in data-driven frameworks, particularly machine learning, have further enhanced the utility of voice-based assessments by uncovering complex patterns linked to cognitive states, including normal cognition (NC), mild cognitive impairment (MCI), and dementia (DE).^{3–11} Machine learning models can analyze large datasets of voice samples to detect subtle changes, providing an objective and scalable approach to cognitive assessment.^{12–14} This technology holds potential for early screening in clinical settings as well as remote monitoring, which could be especially valuable in resource-limited environments. Consequently, voice-based diagnostics are emerging as a promising complement to traditional assessments like neuropsychological tests and neuroimaging.

The use of voice data as an assessment tool for neurodegenerative diseases like Alzheimer's is increasingly relevant in today's aging population, where early detection of cognitive decline can improve patient outcomes through timely interventions.¹⁵ However, voice data introduces privacy challenges due to the personally identifiable information (PII) embedded in recordings, such as gender, accent, and emotional state, as well as subtler speech characteristics that can uniquely identify individuals. These risks are amplified when voice data are processed by automated systems,¹⁶ raising concerns about re-identification and potential misuse of data. Consequently, there is a growing demand for privacy-preserving techniques that protect speaker identity without compromising the diagnostic utility of the data. Existing anonymization methods, such as voice scrambling or noise addition, primarily aim to mask speaker identity but often fail to preserve the features critical for cognitive assessment. Advanced

machine learning techniques, including voice conversion,¹⁷ and adversarial learning,¹⁸ have been developed to address these issues by altering or separating identity-related features from task-relevant ones. Differential privacy methods have also been explored,^{19–21} introducing controlled noise to provide formal privacy guarantees, while feature-level obfuscation techniques, such as perturbing or sanitizing embeddings,²² show promise in balancing privacy and utility. Despite these advancements, achieving an optimal trade-off between privacy preservation and diagnostic accuracy remains a challenge, especially in sensitive domains like cognitive assessments, where preserving discriminative features is important. While numerous studies have focused on privacy protection in speech data, few have specifically addressed voice obfuscation in the context of cognitive assessments. One recent study proposed a privacy-preserving framework for dementia detection, using prosody-based disentanglement of speaker embeddings to obscure speaker identity while maintaining diagnostic accuracy.²³ Such approaches, which effectively address both privacy concerns and the preservation of cognitive features, have the potential to advance the use of voice data in diagnostic applications.

In this study, we developed a computational framework to evaluate how obfuscated voice features support cognitive assessment (Figure 1). Using techniques such as pitch-shifting for voice obfuscation as an example, we analyzed the extent to which derived acoustic features preserved their utility for assessment of cognitive status. This framework was applied to voice data from the Framingham Heart Study (FHS) and the DementiaBank Delaware (DBD) corpus, both consisting of spoken responses to neuropsychological tests. We employed six classification algorithms to evaluate the diagnostic utility of speech features post-obfuscation. To balance speaker privacy and cognitive feature utility, we implemented a weighted linear interpolation approach, which allowed us to adjust the balance between the extent of obfuscation and the preservation of cognitive assessment features. By addressing the dual challenge of privacy risks and diagnostic needs, our framework represents a step forward in enabling secure and effective use of voice data for cognitive assessments.

2 | METHODS

2.1 | Study population

We obtained voice recordings from the FHS and the DBD corpus (Table 1).^{8,9,24–27} FHS is a community-based longitudinal observational study initiated in 1948, which has provided insights into the epidemiology of cardiovascular disease and its risk factors across multiple generations.²⁶ In 1999, FHS initiated a series of investigations into brain structure and cognitive function, recruiting participants for brain MRI scans and neuropsychological testing. The FHS dataset encompasses a variety of neuropsychological tests that assess multiple cognitive domains, including memory, attention, executive function, language, reasoning, visuo perceptual skills, and premorbid intelligence. The cognitive status of participants, classified as NC, MCI, or DE, was determined over time by the FHS dementia diagnosis review panel. For each participant, cognitive status was assigned based on the diagnosis date closest to the recording date, either on or before the recording date, or within 180 days thereafter. DementiaBank,²⁷ part of the TalkBank project, is an open-access repository for multimedia spoken language data aimed at advancing research on language and cognition in dementia. The newly established DBD corpus builds on this initiative by collecting standardized discourse data from NC adults and those with MCI. Participants completed a 90-min session via Zoom, which included a discourse protocol eliciting four types of speech: picture description, story narrative, procedural discourse, and personal narrative, along with a cognitive-linguistic battery. The dataset, available on the DementiaBank website, includes CHAT-formatted transcripts linked to audio recordings, demographic data, and test results.

The FHS speech recordings were obtained in WAV format, with an average duration of 74.36 ± 26.62 min and a sampling rate of 22,050 Hz. The DBD recordings, initially in MP3 format with an average duration of 10.81 ± 4.82 min, were converted to WAV format at a sampling rate of 22,050 Hz to ensure consistency in processing. Both datasets included recordings and transcripts of interactions between examiners and participants, encompassing both questions and responses. Diarization was then applied to isolate participant speech by removing examiner interactions, ensuring the analysis focused on acoustic features relevant to cognitive impairment. The FHS dataset comprised 128 speech samples from participants categorized as NC, MCI, or DE. As of July 2, 2024, data collection for the DBD dataset is ongoing; thus far, it included 85 speech samples from participants with either NC or MCI.

2.2 | Derivation of acoustic features

We extracted acoustic features from voice recordings on both cohorts using the Python *librosa* package.²⁸ A total of 12 distinct sets of features were derived, including statistical measures such as minimum, maximum, mean, standard deviation, and median, resulting in 481 features. The extracted features included parameters like ampli-

RESEARCH IN CONTEXT

- 1. Systematic review:** Voice-based assessments have emerged as a non-invasive method for detecting cognitive decline, leveraging acoustic and linguistic markers such as speech rate, articulation, and pitch variation to differentiate between normal cognition (NC), mild cognitive impairment (MCI), and dementia (DE). Despite their diagnostic potential, voice data carries inherent privacy risks due to identifiable characteristics embedded in speech. Traditional anonymization methods, including voice scrambling and noise addition, often compromise key cognitive markers, limiting their effectiveness in clinical applications.
- 2. Interpretation:** This study introduces a computational framework that applies pitch-shifting to protect speaker identity while preserving acoustic features essential for cognitive assessment. Using data from the Framingham Heart Study (FHS) and DementiaBank Delaware (DBD) corpus, pitch-shifting levels were identified that balance speaker obfuscation (measured by equal error rate [EER]) with diagnostic utility (classification accuracy). By safeguarding privacy without degrading cognitive markers, this approach provides a solution for voice-based diagnostics.
- 3. Future directions:** The findings demonstrate the feasibility of pitch-shifting as a potential privacy-preserving strategy for voice-based cognitive assessments. Future research should focus on adaptive privacy techniques to meet diverse privacy requirements, validate the framework across additional datasets and demographic groups, and explore its integration into broader digital health platforms to ensure secure applications.

tude, root mean square (RMS), spectral coefficients, bandwidth, centroid, flatness, roll-off frequency, zero-crossing rate, tempo, Chroma Energy Normalized Statistics (CENS), and Mel-Frequency Cepstral Coefficients (MFCC) with delta features.²⁹ The MFCC delta features were calculated using Savitzky-Golay filtering to capture temporal derivatives.³⁰

The extracted acoustic features encompass a range of measures that capture various characteristics of the audio signals. Amplitude reflects the signal's loudness or intensity, while the RMS provides a quantitative measure of the signal's overall power. Spectral coefficients represent the frequency distribution, offering insights into the signal's harmonic structure. The bandwidth measures the range of frequencies present, and the spectral centroid indicates the "center of mass" of the frequency spectrum, which corresponds to the perceived brightness of the signal. Spectral flatness quantifies the degree to which the spectrum resembles noise (flat) versus a tonal signal (peaked), and the

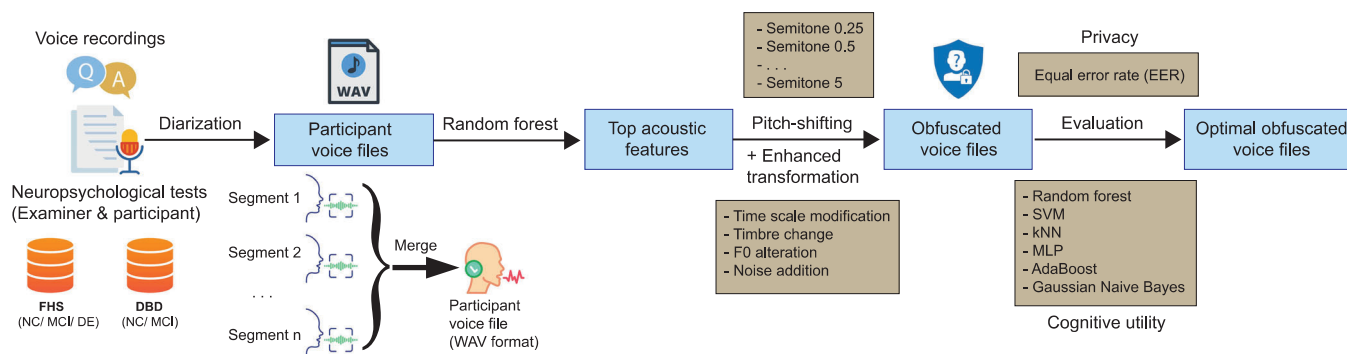


FIGURE 1 Pitch-shifting obfuscation framework for balancing privacy and diagnostic utility. This framework applies pitch-shifting and additional transformations (e.g., time-scale modification, timbre change, F0 alteration, and noise addition) to anonymize speaker identity while preserving cognitive features critical for diagnostic utility. Neuropsychological voice recordings from the Framingham Heart Study (FHS) and DementiaBank Delaware (DBD) corpus were processed, involving diarization to isolate participant speech from examiner dialogue, followed by merging participant-specific voice data and extracting top acoustic features using a random forest model. Obfuscated voice files were assessed for privacy (via equal error rate, EER) and diagnostic utility (using six classification algorithms: Random Forest, Support Vector Machine [SVM], k-Nearest Neighbors [kNN], Multi-Layer Perceptron [MLP], adaptive boosting (AdaBoost), and Gaussian Naive Bayes) across three cognitive states: normal cognition (NC), mild cognitive impairment (MCI), and dementia (DE).

TABLE 1 Study population

Dataset	Cognitive status	Male percentage (%)	Age (mean \pm std)	Speech duration (mean \pm std minutes)
FHS (N = 128)	NC (N = 40)	Male = 60 (47%)	82.84 \pm 8.22	74.36 \pm 26.62
	MCI (N = 10)	NC = 25 (19%)	NC = 80.37 \pm 10.46	NC = 75.05 \pm 25.3
	DE (N = 78)	MCI = 6 (5%)	MCI = 80.3 \pm 7.63	MCI = 57.83 \pm 24.08
		DE = 29 (23%)	DE = 84.43 \pm 6.39	DE = 76.13 \pm 26.87
Dementia Bank (N = 85)	NC (N = 34)	Male = 27 (32%)	71.36 \pm 7.38	10.81 \pm 4.82
	MCI (N = 51)	NC = 5 (6%)	NC = 68.23 \pm 5.82	NC = 12.1 \pm 5.7
		MCI = 22 (26%)	MCI = 73.45 \pm 7.57	MCI = 9.94 \pm 3.91

Note: Demographic and cognitive status characteristics of the Framingham Heart Study (FHS) and DementiaBank Delaware (DBD) datasets. The table provides information on the number of participants (N), cognitive status classification (normal cognition [NC], mild cognitive impairment [MCI], dementia [DE]), percentage of male participants, average age with standard deviation (mean \pm std), and average speech duration in minutes (mean \pm std) for both datasets.

roll-off frequency defines the point below which a specified percentage of spectral energy is concentrated. The zero-crossing rate measures how frequently the signal waveform crosses the zero-amplitude axis, often correlating with signal sharpness or noisiness. Tempo refers to the perceived speed or rhythm of the audio. CENS capture pitch and harmonic content across the chromatic scale, offering insights into melodic and harmonic patterns. Finally, MFCCs and their derivatives (delta features) model the spectral envelope in a perceptually meaningful way, making them particularly valuable for characterizing timbre and speech-related features.

2.3 | Evaluation of privacy and cognitive integrity

Our framework is designed to balance speaker privacy with cognitive diagnostic utility by leveraging acoustic feature extraction, obfusca-

tion methods, and evaluation metrics. To achieve speaker obfuscation, we applied pitch-shifting at levels ranging from 0.25 to 5 semitones and incorporated additional transformations, such as time-scale modifications and noise addition, to alter vocal characteristics. Privacy was quantified using the equal error rate (EER) from an automatic speaker verification (ASV) system,³¹ while diagnostic utility was measured by the classification accuracy (ACC) of machine learning models used to differentiate cognitive states.

To systematically evaluate the effects of pitch-shifting, we normalized pitch shifts on a scale from 0.05 to 1, creating 20 configurations. For each configuration, we calculated classification accuracy across six machine learning algorithms: Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Multi-Layer Perceptron (MLP), adaptive boosting (AdaBoost), and Gaussian Naive Bayes. Further, statistical analyses were conducted on 20 distinct EER and ACC values to compute the mean, standard deviation, and corresponding *p*-values.

Data normality was assessed using the Shapiro-Wilk test. For normally distributed data, a one-sample t-test was performed, while the Wilcoxon signed-rank test was applied to non-normally distributed data to determine statistical significance. This approach allowed us to assess how acoustic transformations impacted the ability to classify cognitive states accurately. Privacy degradation was evaluated by calculating the EER for original versus obfuscated voice files, providing a robust measure of speaker identity protection. To explore the trade-off between privacy and diagnostic utility, we combined EER and ACC using a weighted linear interpolation approach, quantified by an overall performance metric defined as

$$\rho = \alpha (EER) + (1 - \alpha) (ACC), \quad (1)$$

where α represents the privacy level ranging from 0 to 1. Setting $\alpha = 0.5$ gave equal importance to privacy and utility.

2.4 | Additional transformations to enhance privacy

In addition to pitch-shifting, we explored a range of acoustic transformations to enhance speaker obfuscation. Time-scale modification was employed to adjust the speed of the audio signal, which introduced temporal variability to the vocal characteristics. Timbre change was implemented to modify the harmonic quality of the voice, effectively altering its tonal color and making speaker identification more challenging. Fundamental frequency (F0) alteration was utilized to shift the natural pitch contour of the voice, introducing variations in vocal inflection that further masked speaker identity. Finally, controlled noise addition was applied to introduce subtle distortions in the audio signal, reducing the fidelity of speaker-specific acoustic features while retaining critical components needed for cognitive assessment.

3 | RESULTS

3.1 | Acoustic features for cognitive assessment

The FHS dataset had a longer average speech duration (74.36 ± 26.62 min) compared to the DBD dataset (10.81 ± 4.82 min). Longer samples provided a broader range of acoustic features and richer vocal characteristics for analysis. To identify the most important acoustic features relevant for cognitive assessment, supervised learning techniques were employed, with each participant's speech data labeled according to their cognitive status: NC, MCI, or DE. A random forest algorithm, utilizing an ensemble of 100 decision trees, was applied to classify the recordings based on cognitive status and assess the importance of each feature for predictive performance. The resulting importance scores were normalized between 0 and 1, and ranked.

The top 20 acoustic features with the highest importance scores were selected from both datasets for further analysis (Data S1 and S2). In the FHS dataset, key features included statistical measures

of MFCC, MFCC delta, zero-crossing rate, CENS, and spectral bandwidth. Similarly, in the DBD dataset, prominent features comprised statistical measures of MFCC, MFCC delta, and CENS. To evaluate the impact of feature selection on classification accuracy, the top-k features were iteratively included (ranging from $k = 1$ to $k = 481$), and six classification algorithms: Random Forest, SVM, kNN, MLP, AdaBoost, and Gaussian Naive Bayes, were applied using 10-fold cross-validation (Data S3 and S4). The classification tasks involved distinguishing between NC, MCI, and DE for the FHS dataset and between NC and MCI for the DBD dataset.

The performance trends revealed Random Forest as the most robust algorithm across varying feature dimensions, achieving an accuracy of approximately 70% using 20 features in the FHS dataset. SVM showed similar performance, whereas kNN and MLP exhibited decreased accuracy and higher fluctuation with additional features (Figure 2A). For the DBD dataset, Random Forest and Gaussian Naive Bayes demonstrated minimal accuracy variation with an increasing number of features, whereas other algorithms experienced a decline in accuracy (Figure 2B). Reducing the feature set to the top 20 features improved average classification accuracy by 0.0149 for the FHS dataset and 0.1523 for the DBD dataset across all algorithms (Table S1). This demonstrates that a reduced feature set not only maintains classification performance but also enhances computational efficiency. Additionally, employing Random Forest with the top-ranked features yielded improved area under the curve (AUC) scores for cognitive state classification tasks. For instance, the NC/MCI binary classification task in the DBD dataset showed an AUC improvement of up to 0.2 compared to using the full feature set (Figures 3A,B). These results highlight the potential of selecting a focused set of high-impact acoustic features to improve diagnostic utility, computational efficiency, and overall model performance in voice-based cognitive assessments.

3.2 | Balancing privacy and diagnostic utility

Our results demonstrate a trade-off analysis between privacy (quantified by EER) and diagnostic utility (measured by classification accuracy) for FHS and DBD (Figure 4). In both datasets, the optimal balance between privacy and utility was determined through weighted linear interpolation using a total performance metric (ρ). The FHS dataset achieved a peak ρ of 0.478 at a pitch-shifting level of 3.25 semitones, while the DBD dataset attained a maximum ρ of 0.408 at a pitch-shifting level of 2.75 semitones.

The two plots illustrate the relationship between classification accuracy (utility) and privacy levels (denoted by EER) for the FHS and DBD datasets. Privacy levels are manipulated using pitch-shifting transformations applied to voice recordings, with higher values of EER corresponding to greater obfuscation of speaker identity. In the FHS dataset (Figure 4A), the red line represents the trend of classification accuracy as a function of privacy level, while the shaded region indicates the 95% confidence interval. The mean and standard deviation of EER and classification accuracy are 0.33 ± 0.013 (p -value = 4.38×10^{-28}) and 0.624 ± 0.017 (p -value = 4.38×10^{-28}), respectively. The initial

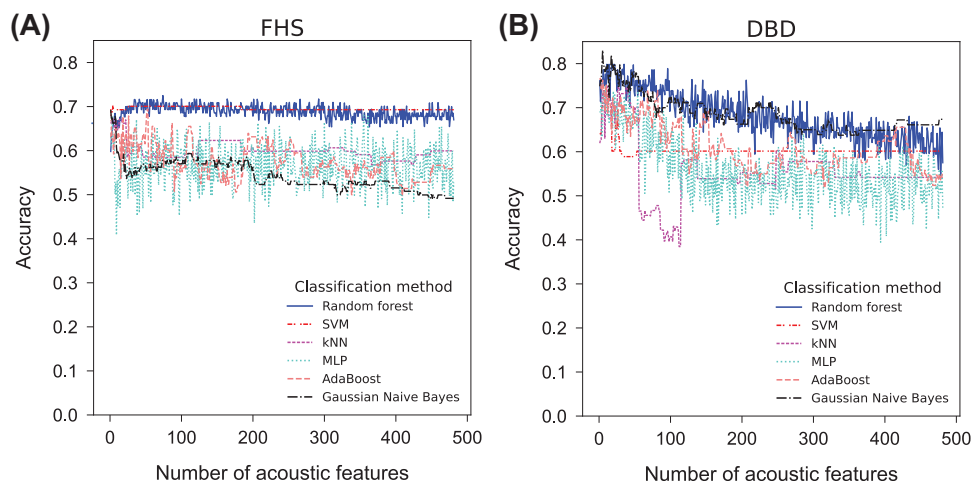


FIGURE 2 Impact of acoustic feature count on model accuracy. Classification accuracy as a function of the number of acoustic features for six machine learning algorithms (Random Forest, Support Vector Machine (SVM), k-Nearest Neighbors (kNN), Multi-Layer Perceptron (MLP), adaptive boosting (AdaBoost), and Gaussian Naive Bayes) applied to digital voice recordings from the (A) Framingham Heart Study (FHS) and (B) DementiaBank Delaware (DBD) corpus.

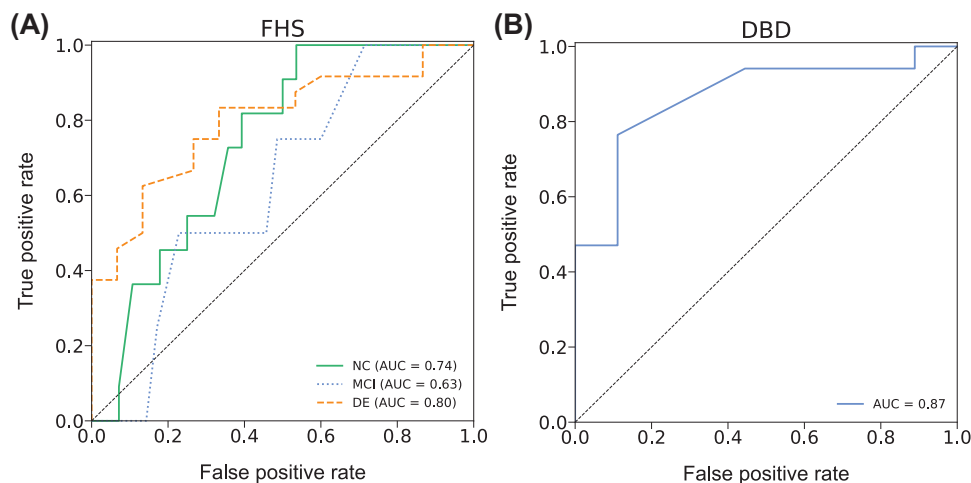


FIGURE 3 Model performance. Receiver operating characteristics (ROC) curves demonstrating classification performance of models trained on the top 20 acoustic features for classification of cognitive status in (A) the Framingham Heart Study (FHS) cohort and (B) the Dementia Bank Delaware (DBD) corpus. (A) Area under the curve (AUC) values for distinguishing normal cognition (NC), mild cognitive impairment (MCI), and dementia (DE) in the FHS cohort. (B) An overall AUC for distinguishing NC from MCI cases in the DBD cohort.

accuracy is approximately 0.646, which gradually declines to around 0.598 at the maximum privacy level. This trend demonstrates the trade-off between privacy preservation and diagnostic utility, where increasing privacy reduces classification accuracy. The spread of data points around the trendline reflects variations in model performance, likely influenced by different feature combinations and classifiers. For the DBD dataset (Figure 4B), the red line similarly depicts the trend of classification accuracy across privacy levels, with a more pronounced negative slope compared to the FHS dataset. The mean and standard deviation of EER and classification accuracy are 0.157 ± 0.115 (p -value = 1.91×10^{-6}) and 0.641 ± 0.049 (p -value = 7.52×10^{-6}), respectively. The accuracy begins at approximately 0.722 and steadily decreases to

around 0.572 as privacy levels increase. This sharper decline suggests that pitch-shifting has a stronger impact on cognitive feature preservation in shorter-duration recordings, which characterize the DBD dataset. The shaded confidence interval highlights variability, with larger fluctuations observed at higher privacy levels. Together, these plots highlight the inherent trade-off in balancing privacy and utility in voice-based cognitive assessments. Lower privacy levels retain more diagnostic utility, while higher privacy levels compromise accuracy.

To enhance the privacy capabilities of our framework, four additional transformation layers, time-scale modification (scale: 1.2), timbre alteration (scale: 1.1–1.3), fundamental frequency (F0) adjustment (scale: 1.1–1.3), and noise addition (scale: 0.001), were applied to

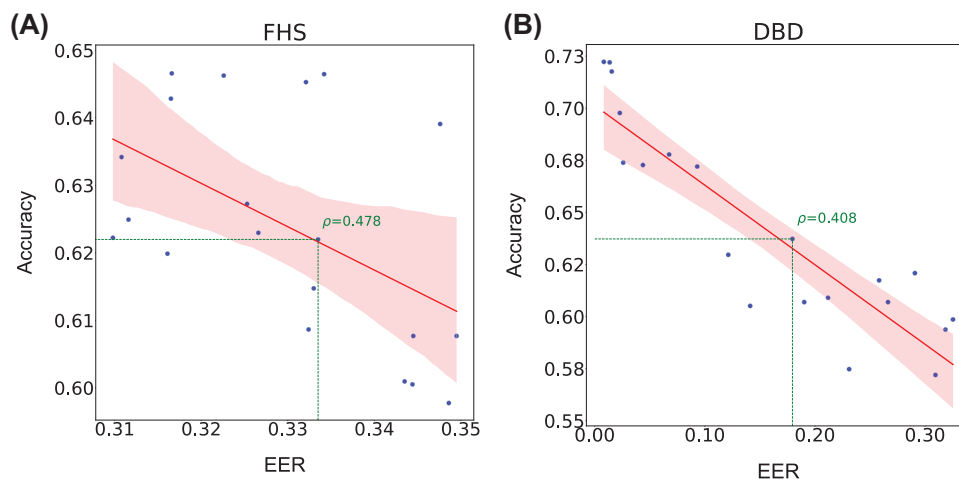


FIGURE 4 Privacy-utility trade-off analysis. (A) Regression plot displaying the relationship between privacy, measured as equal error rate (EER), and utility, measured as classification accuracy, for the Framingham Heart Study (FHS) dataset. Model performance was evaluated across 20 distinct pitch-shifting levels (0.25 to 5 semitones), with a 95% confidence interval shown. The green dashed lines indicate a trade-off point, achieving an EER of 0.3335 and an accuracy of 0.6220, corresponding to a total performance score of 0.478 at $\alpha = 0.5$. This balance was achieved at a pitch-shifting level of 3.25 semitones. (B) Similar regression plot for the DementiaBank Delaware (DBD) dataset, illustrating the relationship between EER and classification accuracy across the same 20 pitch-shifting levels. The green dashed lines identify a trade-off, with an EER of 0.1796 and an accuracy of 0.6372, resulting in a total performance score of 0.408 at $\alpha = 0.5$. The optimal trade-off corresponds to a pitch-shifting level of 2.75 semitones.

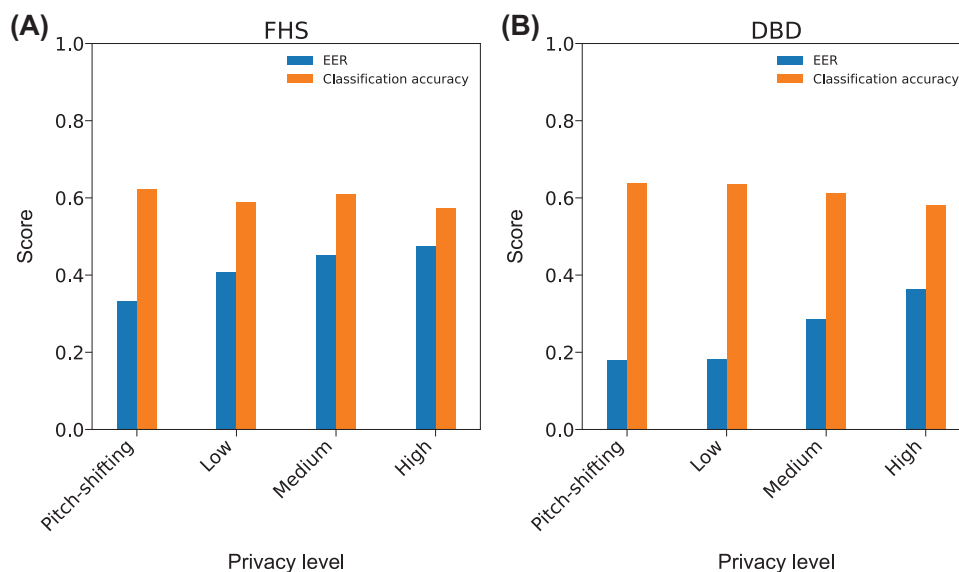


FIGURE 5 Privacy-utility trade-off with additional enhancements. Comparison of standard pitch-shifting and an enhanced approach incorporating additional transformations, including time-scale modification (scale: 1.2), timbre alteration (scale: 1.1–1.3), fundamental frequency (F0) adjustment (scale: 1.1–1.3), and small noise addition (scale: 0.001). Performance was evaluated across three transformation levels: low, medium, and high. (A) Privacy-utility analysis on the Framingham Heart Study (FHS) dataset, showing changes in equal error rate (EER) and classification accuracy as transformation intensity increases. (B) Privacy-utility analysis on the DementiaBank Delaware (DBD) dataset, illustrating the impact of varying transformation levels on EER and classification accuracy.

the optimally pitch-shifted voice files. These transformations were implemented incrementally at low, medium, and high levels. Results showed that adding these transformations improved privacy while reducing classification accuracy, reflecting the privacy-utility trade-off (Figures 5A,B). For the FHS dataset, privacy enhancement was

observed as the EER score increased from 0.3335 (optimal pitch-shifting level of 3.25 semitones) to 0.4746 at the high transformation level. However, this improvement in privacy came at a cost, with classification accuracy decreasing from 0.6220 to 0.5725 (Data S5). Similarly, in the DBD dataset, the EER score improved from 0.1796 (optimal

pitch-shifting level of 2.75 semitones) to 0.3630 at the high transformation level, while classification accuracy declined from 0.6372 to 0.5812 (Data S6). These results underscore a significant trade-off: while additional transformation layers enhance privacy by making speaker obfuscation more robust, they also diminish the diagnostic utility of the framework in distinguishing between cognitive states. This trade-off must be carefully managed depending on the specific application requirements, balancing privacy needs against diagnostic accuracy.

4 | DISCUSSION

Our findings highlight the intricate trade-off between preserving speaker privacy and maintaining the diagnostic utility of acoustic features for cognitive assessments. The differences in optimal pitch-shifting levels observed between the FHS and DBD datasets reflect these dataset-specific factors. The longer average speech duration in the FHS dataset enabled the retention of a richer range of acoustic features, resulting in a more balanced trade-off between privacy and utility. Conversely, the shorter average duration of recordings in the DBD dataset required more aggressive pitch-shifting to achieve comparable levels of privacy, albeit at the expense of classification accuracy. These results emphasize the importance of considering dataset-specific factors when applying the proposed framework, particularly when translating it to diverse real-world settings. The implications of this work extend beyond the immediate context of dementia diagnostics. For instance, as voice data becomes increasingly integral to telemedicine and remote patient monitoring, these techniques could be adapted to protect patient identities while enabling meaningful clinical insights. Moreover, this study underscores the importance of dataset-specific considerations, suggesting that generalizable privacy-utility frameworks must account for varying data characteristics, such as recording length and feature distribution.

Our study has a few limitations. Using pitch-shifting as an obfuscation method inherently involves trade-offs, as vocal frequency is a critical feature in speech-based cognitive assessment. While our primary aim was not to advocate for pitch-shifting as the optimal obfuscation method, it served as a practical example to demonstrate the utility of our framework in balancing privacy preservation and diagnostic utility. We acknowledge that altering vocal frequency modifies aspects of the original speech signal, potentially impacting cognitive assessments. This limitation highlights the broader challenge faced by obfuscation techniques, as any such method is likely to alter vocal features. To address this, our weighted linear interpolation framework provides a systematic approach to adjust the balance between privacy and diagnostic utility, allowing flexibility to explore varying levels of obfuscation tailored to specific applications. Our study also focused on a limited set of acoustic features and classification algorithms. Future work should explore a broader range of features, including temporal and prosodic elements, and additional machine learning methods to enhance classification accuracy across diverse populations.

Expanding the research to more diverse cohorts would strengthen the generalizability and clinical applicability of pitch-shifting techniques. Another limitation is the potential for re-identification in white-box scenarios, where an attacker familiar with the pitch-shifting method could reverse-engineer the obfuscation. Exploring adaptive models to counteract such attacks could improve security.

In conclusion, our study introduces a computational framework for balancing privacy preservation and diagnostic utility in voice-based cognitive assessments. By leveraging techniques such as pitch-shifting as means of voice obfuscation, we demonstrated the ability to mitigate privacy risks while preserving the diagnostic value of acoustic features. Using weighted linear interpolation, our approach identifies optimal trade-offs, setting the stage for future exploration of more advanced obfuscation methods in digital health applications. This work contributes to the ethical and practical integration of voice data in medical analyses, emphasizing the importance of protecting patient privacy while maintaining the integrity of cognitive health assessments. These findings pave the way for developing standardized, privacy-centric guidelines for future applications of voice-based assessments in clinical and research settings.

ACKNOWLEDGMENTS

This project was supported by grants from the National Institute on Aging's Artificial Intelligence and Technology Collaboratories (P30-AG073104 and P30-AG073105), the American Heart Association (20SFRN35460031), Gates Ventures, and the National Institutes of Health (R01-HL159620, R01-AG062109, and R01-AG083735).

CONFLICT OF INTEREST STATEMENT

V.B.K. is a co-founder and equity holder of deepPath Inc. and CogniScreen, Inc. He also serves on the scientific advisory board of Altoida Inc. R.A. is a scientific advisor to Signant Health and NovoNordisk. The remaining authors declare no competing interests. Author disclosures are available in the [Supporting Information](#).

CONSENT STATEMENT

All participants included in this study provided informed consent, with the understanding that their speech data would be used for research purposes related to cognitive assessment. This study has been reviewed and approved by the relevant institutional ethics committee, confirming that it adheres to guidelines for responsible data handling and participant privacy.

ORCID

Vijaya B. Kolachalama  <https://orcid.org/0000-0002-5312-8644>

REFERENCES

1. Mahon E, Lachman ME. Voice biomarkers as indicators of cognitive changes in middle and later adulthood. *Neurobiol Aging*. 2022;119:22-35.
2. Ginsberg SD, Themistocleous C, Eckerström M, Kokkinakis D. Voice quality and speech fluency distinguish individuals with Mild Cognitive Impairment from Healthy Controls. *PLoS One*. 2020;15:e0236009.

3. Amini S, Hao B, Zhang L, et al. Automated detection of mild cognitive impairment and dementia from voice recordings: a natural language processing approach. *Alzheimers Dement*. 2022;19(3):946-955.
4. Chen J, Ye J, Tang F, Zhou J. Automatic detection of alzheimer's disease using spontaneous speech only. *Interspeech*. 2021;2021:3830-3834.
5. Ding H, Lister A, Karjadi C, et al. Detection of mild cognitive impairment from non-semantic, acoustic voice features: the Framingham heart study. *JMIR Aging*. 2024;7:e55126.
6. Haider F, de la Fuente S, Luz S. An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE J Sel Top Signal Process*. 2020;14:272-281.
7. Hajjar I, Okafor M, Choi JD, et al. Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early Alzheimer's disease. *Alzheimers Dement (Amst)*. 2023;15:e12393.
8. Karjadi C, Xue C, Cordella C, et al. Fusion of low-level descriptors of digital voice recordings for dementia assessment. *J Alzheimers Dis*. 2023;96:507-514.
9. Xue C, Karjadi C, Paschalidis IC, Au R, Kolachalama VB. Detection of dementia on voice recordings using deep learning: a Framingham Heart Study. *Alzheimers Res Ther*. 2021;13:146.
10. Lin K, Washington PY. Multimodal deep learning for dementia classification using text and audio. *Sci Rep*. 2024;14:13887.
11. Nishikawa K, Akihiro K, Hirakawa R, Kawano H, Nakatoh Y. Machine learning model for discrimination of mild dementia patients using acoustic features. *Cogn Robot*. 2022;2:21-29.
12. Shi M, Cheung G, Shahamiri SR. Speech and language processing with deep learning for dementia diagnosis: a systematic review. *Psychiatry Res*. 2023;329:115538.
13. Yang Q, Li X, Ding X, Xu F, Ling Z. Deep learning-based speech analysis for Alzheimer's disease detection: a literature review. *Alzheimer's Res Ther*. 2022;14:186.
14. Javeed A, Dallora AL, Berglund JS, Ali A, Ali L, Anderberg P. Machine learning for dementia prediction: a systematic review and future research directions. *J Med Syst*. 2023;47:17.
15. Pulido MLB, Hernández JBA, Ballester MÁF, González CMT, Mekyska J, Smékal Z. Alzheimer's disease and automatic speech analysis: a review. *Expert Syst Appl*. 2020;150:113213.
16. Nautsch A, Jiménez A, Treiber A, et al. Preserving privacy in speaker and speech characterisation. *Comput Speech Lang*. 2019;58:441-480.
17. Huang F, Zeng K, Zhu W. DiffVC+: improving Diffusion-based Voice Conversion for Speaker Anonymization. *Interspeech*. 2024;2024:4453-4457.
18. Yang S, Tantrawenith M, Zhuang H, et al. Speech Representation Disentanglement with Adversarial Mutual Information Learning for One-shot Voice Conversion. *Interspeech*. 2022;2022:2553-2557.
19. Tomashenko N, Wang X, Vincent E, et al. The VoicePrivacy 2020 Challenge: results and findings. *Comput Speech Lang*. 2022;74:101362.
20. Aloufi R, Haddadi H, Boyle D. Privacy-preserving Voice Analysis via Disentangled Representations. Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop 2020. p. 1-14.
21. Abadi M, Chu A, Goodfellow I, et al. Deep Learning with Differential Privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security 2016. p. 308-318.
22. Lal Srivastava BM, Vauquier N, Sahidullah M, Bellet A, Tommasi M, Vincent E. Evaluating Voice Conversion-Based Privacy Protection against Informed Attackers. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020. p. 2802-2806.
23. Woszczyk D, Aloufi R, Demetriou S. Prosody-driven privacy-preserving dementia detection. *Interspeech*. 2024;2024:3035-3039.
24. Lanzi AM, Saylor AK, Fromm D, MacWhinney B, Cohen ML. Establishing the dementiaBank delaware corpus: an online multimedia database for the study of language and cognition in dementia. *Alzheimer's & Dementia*. 2023;19(S19):e073058.
25. Lanzi AM, Saylor AK, Fromm D, Liu H, MacWhinney B, Cohen ML. DementiaBank: theoretical Rationale, Protocol, and Illustrative Analyses. *Am J Speech-Lang Pathol*. 2023;32:426-438.
26. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*. 2014;383:999-1008.
27. Lanzi AM, Saylor AK, Fromm D, Liu H, MacWhinney B, Cohen ML. DementiaBank: theoretical rationale, protocol, and illustrative analyses. *Am J Speech Lang Pathol*. 2023;32:426-438.
28. McFee B, Raffel C, Liang D, et al. librosa: audio and Music Signal Analysis in Python. Proceedings of the 14th Python in Science Conference 2015. 18-24.
29. Abraham JVT, Khan AN, Shahina A. A deep learning approach for robust speaker identification using chroma energy normalized statistics and mel frequency cepstral coefficients. *Int J Speech Technol*. 2021;26:579-587.
30. Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal Chem*. 2002;36:1627-1639.
31. Jyh-Min C, Hsiao-Chuan W. A method of estimating the equal error rate for automatic speaker verification. SympoTIC '04 Joint 1st Workshop on Mobile Future & Symposium on Trends In Communications (IEEE Cat No04EX877)2004. 285-288.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ahangaran M, Dawalatabad N, Karjadi C, Glass J, Au R, Kolachalama VB. Obfuscation via pitch-shifting for balancing privacy and diagnostic utility in voice-based cognitive assessment. *Alzheimer's Dement*. 2025;21:e70032. <https://doi.org/10.1002/alz.70032>