scientific reports

Check for updates

OPEN

The effect of speech pathology on automatic speaker verification: a large-scale study

Soroosh Tayebi Arasteh[1,2,3]✉, Tobias Weise[1,2], Maria Schuster[4], Elmar Noeth[1], Andreas Maier[1] & Seung Hee Yang[2]

Navigating the challenges of data-driven speech processing, one of the primary hurdles is accessing reliable pathological speech data. While public datasets appear to offer solutions, they come with inherent risks of potential unintended exposure of patient health information via re-identification attacks. Using a comprehensive real-world pathological speech corpus, with over n=3800 test subjects spanning various age groups and speech disorders, we employed a deep-learning-driven automatic speaker verification (ASV) approach. This resulted in a notable mean equal error rate (EER) of 0.89 ± 0.06%, outstripping traditional benchmarks. Our comprehensive assessments demonstrate that pathological speech overall faces heightened privacy breach risks compared to healthy speech. Specifically, adults with dysphonia are at heightened re-identification risks, whereas conditions like dysarthria yield results comparable to those of healthy speakers. Crucially, speech intelligibility does not influence the ASV system's performance metrics. In pediatric cases, particularly those with cleft lip and palate, the recording environment plays a decisive role in re-identification. Merging data across pathological types led to a marked EER decrease, suggesting the potential benefits of pathological diversity in ASV, accompanied by a logarithmic boost in ASV effectiveness. In essence, this research sheds light on the dynamics between pathological speech and speaker verification, emphasizing its crucial role in safeguarding patient confidentiality in our increasingly digitized healthcare era.

Background

Speech is a biomarker that is extensively explored for the development of healthcare applications because of its low cost and non-invasiveness[1]. With the advances in deep learning (DL), data-driven methods have gained a lot of attention in speech processing in healthcare[2]. For example, in the medical domain, speech biomarker reflects objective measurement that can be used for accurate and reproducible diagnosis. From diagnosis[3–6] to therapy[7–9], pathological speech could be a rich source for different data-driven applications in healthcare. This is critical to the rapid and reliable development of medical screening, diagnostics, and therapeutics. However, accessing pathological speech data for utilization in computer-assisted methods is a challenging and time-consuming process because of patient privacy concerns leading to the fact that most studies only investigated small cohorts due to the resulting lack of data[10].

Related works

Pathological speech has garnered significant attention in DL-based automatic analyses of speech and voice disorders. Notably, Vásquez-Correa et al.[11] broadly assessed Parkinson's disease, while Rios-Urrego et al.[12] delved into evaluating the pronunciation skills of Parkinson's disease patients. Such works emphasize the potential of pathological speech as an invaluable resource for Parkinson's disease analysis. Additionally, numerous studies have employed pathological speech for DL-based analyses of Alzheimer's disease. Pérez-Toro et al.[13] illustrated the efficacy of the Arousal Valence plane for discerning and analyzing depression within Alzheimer's disease. Pappagari et al.[4] fused speaker recognition and language processing techniques for assessing the severity of

[1]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany. [2]Speech & Language Processing Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany. [3]Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, 52074 Aachen, Germany. [4]Department of Otorhinolaryngology, Head and Neck Surgery, Ludwig-Maximilians-Universität München, 80333 Munich, Germany. ✉email: soroosh.arasteh@fau.de

Alzheimer's disease. Furthermore, García et al.'s work[14] delved into dysphonia assessment, Kohlschein et al.[15] addressed aphasia, Bhat et al.[16] explored dysarthria, and Gargot et al.[17] investigated Autism Spectrum Disorders.

The burgeoning role of pathological speech in healthcare is evident, especially as computer-assisted, data-driven methods continue to flourish. However, this growth is tempered by the challenges in accessing pathological speech data. Patient privacy concerns make this not only a daunting task but also a protracted endeavor. Within this framework emerges a pivotal question: Does pathological speech, when examined as a biomarker, possess a heightened susceptibility to re-identification attacks compared to healthy speech? Addressing this necessitates the incorporation of ASV-a tool that verifies if an unrecognized voice belongs to a specific individual-to ascertain the privacy levels inherent to healthy speech data[18].

Laying the groundwork for understanding biomarkers in clinical research, Strimbu et al.[19] and Califf et al.[20] have proffered working definitions and established a foundational framework. Delving deeper, Marmar et al.[21] elucidated the diagnostic potential of speech-based markers, particularly in identifying posttraumatic stress disorder, while Ramanarayanan et al.[22] unpacked both the opportunities and the impediments associated with harnessing speech as a clinical biomarker. Remarkably, existing literature remains silent on the interplay between speech pathology and ASV. Our study is thus positioned to fill this void, venturing to discern the relative vulnerability of pathological speech to re-identification in contrast with its healthy counterpart.

Main contributions

In this study, we undertake a detailed look at how pathological speech affects ASV. We use a large and real-world dataset[23] of around 200 hours of recordings that includes both pathological and healthy recordings. Our research focuses on text-independent speaker verification (TISV), to capture a broader range of scenarios[24,25]. Considering the many factors that can sway ASV results, we made efforts to keep various conditions consistent by:

1. Equalizing the training and test set sizes,
2. ensuring consistent sound quality across recordings,
3. matching age distributions within different subgroups,
4. regulating background noise,
5. controlling for the type of microphone utilized and the recording environment, and
6. grouping by specific pathologies.

In the sections that follow, we break down our findings methodically:

- We start with broad-spectrum experiments to paint a comprehensive picture of our ASV system's prowess using the entire pathological dataset.
- Subsequently, our exploration narrows, dissecting the influence of specific pathologies on ASV for both adults and children.
- We then examine how combining data from different speech problems affects ASV. We also look into how the size of the training dataset influences ASV performance.
- Concluding our findings, we assess the influence of speech intelligibility on ASV's performance.

We assume that equal error rate (EER) is a measure of anonymity in the dataset. The lower the EER, the higher the vulnerability of the respective group. This is also a common choice in speaker verification challenges[18]. Furthermore, we use word recognition rate (WRR) as a measure of speech intelligibility as it demonstrated high and significant correlations in many previous studies[10,23,26,27]. The lower the WRR the less intelligible, the speech of the persons in the respective group.

Our goal is to uncover the connection between pathological speech conditions and speaker verification's success rate. We show evidence that the distinct features of pathological speech, when paired with different recording conditions, influence speaker verification outcomes.

Results

Pathology influences ASV performance

When examining pathological recordings from both adult and child subsets, our results showed a mean EER of $0.89 \pm 0.06\%$. For this, n=2064 speakers were used for training and n=517 for testing. Notably, this EER is lower than common values found in datasets such as LibriSpeech[30] or VoxCeleb1 &2[31,32]. This outcome was the average from 20 repeated experiments to counteract the potential biases of random sampling. To ensure an equitable comparison across groups, each subgroup was adjusted in terms of age distribution and speaker numbers. After employing standard training and evaluation, we then evaluated the speaker verification outcomes for each subgroup against control groups.

Adults: Adult patients were divided into three categories: "dysglossia-dnt", "dysarthria-plant", and "dysphonia-logi". For benchmarking, $n = 85$ healthy individuals formed the control group, labeled as "ctrl-plant-A". When examining EER values, both "dysglossia-dnt-85" ($3.05 \pm 0.74\%$) and "dysarthria-plant-85" ($2.91 \pm 1.09\%$) showed no significant difference from the control group "ctrl-plant-A-85" ($3.12 \pm 0.94\%$) with P= 0.786 and P= 0.520, respectively. In contrast, the "dysphonia-logi-85" group, at an EER of $2.40 \pm 0.84\%$, was significantly different from the control, with a P= 0.015. Refer to Fig. 1 for a visual representation of these findings.

Children: Children were divided into two categories, "CLP-dnt" and "CLP-plant", both representing patients with cleft lip and palate (CLP). Additionally, a control group of $n = 1662$ healthy children, "ctrl-plant-C", was used. As for the EER values, "CLP-dnt-124" yielded ($5.25 \pm 0.90\%$, which was not significantly different from the

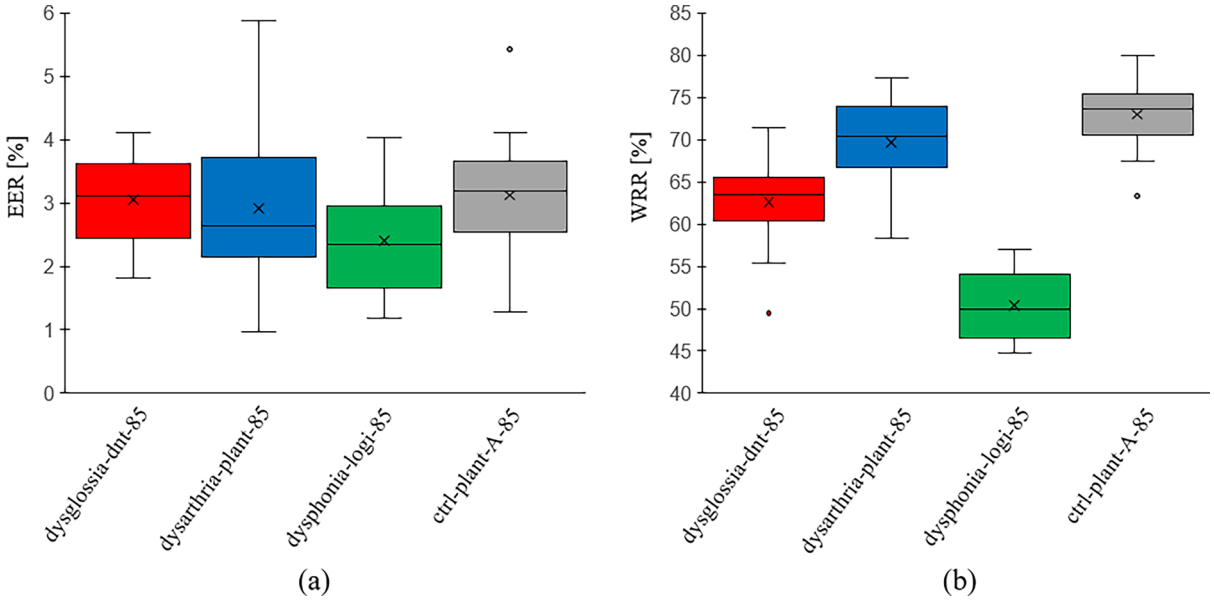

Figure 1. Evaluation results of speaker verification on the adults for individual groups for 20 repetitions. During each repetition, n=85 speakers are sampled for each group and n=68 of them were assigned to training and n=17 speakers to test. All the values are given in percent. (**a**) Equal error rate (EER) values. (**b**) Word recognition rate (WRR) values. *dysglossia* Patients with dysglossia who underwent prior maxillofacial surgery, *dysarthria* Patients diagnosed with dysarthria, *dysphonia* Patients with voice disorders, *CLP* children with cleft lip and palate, *dnt* recordings from the "dnt Call 4U Comfort" headset[27], *plant* recordings via Plantronics Inc. headset[28], *logi:* recordings via Logitech International S.A. headset[29], *ctrl* control group. Numbers appended, such as "-85" in "dysglossia-dnt-85", represent the total speaker count for that experiment.

control group's 5.72 ± 1.05 (P= 0.134). However, "CLP-plant-124" stood out at 7.82 ± 0.91%, differing significantly from the control's 5.72 ± 1.05% (P< 0.001). Figure 2 offers a detailed visual comparison among the groups.

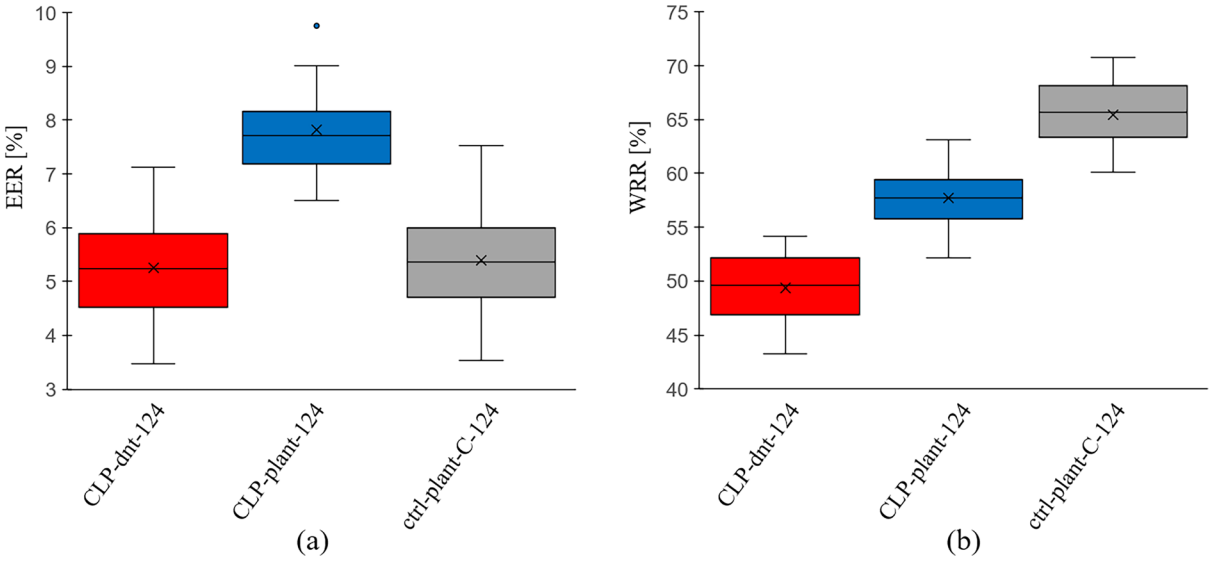

Figure 2. Evaluation results of speaker verification on the children for individual groups for 20 repetitions. During each repetition, 124 speakers are sampled for each group and 99 of them were assigned to training and 25 speakers to test. All the values are given in percent. (**a**) Equal error rate (EER) values. (**b**) Word recognition rate (WRR) values. *CLP* children with cleft lip and palate, *dnt* recordings from the "dnt Call 4U Comfort" headset[27], *plant* recordings via Plantronics Inc. headset[28], *ctrl* control group. Numbers appended, such as "-124" in "CLP-dnt-124", represent the total speaker count for that experiment.

Pathological diversity in speakers leads to substantial reduction in ASV error rate

In our pursuit to understand the influence of pathological diversity on ASV, various datasets were combined, maintaining the speaker count for training and testing as for the children (see Table 2), with an age distribution to match. Upon combining the variations from the "all-children-124" set, we noticed a notable improvement in average EER. Specifically, it stood at 4.80 ± 0.98%, which was considerably better than the control group "ctrl-plant-C-124" that recorded an EER of 5.72 ± 1.05% (P= 0.006). This data highlights the potential benefits of integrating multiple sources of variation in reducing error rates.

Further, when leveraging larger training sets infused with pathological diversity (see Table 2), the EER for the mixed pathological group "CLP-dnt-plant-500" was 2.88 ± 0.25%. In comparison, the EER for the healthy group "ctrl-plant-C-500" was 3.04 ± 0.17% (P= 0.020). This reinforces the premise that the pathological group, with its inherent diversity, offers an advantage in speaker verification over the relatively homogenous healthy group.

Increase in training speaker number yields logarithmic enhancement in ASV performance

Exploring the impact of training set size on ASV performance, we integrated both pathological and healthy speakers from a comprehensive pool of n=3,849. Various speaker groups were drawn from this collective, and they underwent our standard training and evaluation processes.

The "all-spk-50" dataset, which comprised 50 speakers, recorded an EER of 5.19 ± 1.63%. With an increased speaker count in the "all-spk-500" dataset, the EER was reduced to 1.87 ± 0.19%, marking a significant improvement with a P< 0.001. Extending the dataset to 1500 speakers ("all-spk-1500"), the EER further decreased to1.15 ± 0.10%, surpassing the performance of the previous group with a P< 0.001. When the dataset was expanded to 3,000 speakers ("all-spk-3000"), the EER diminished to 0.90 ± 0.05%, outperforming the 1,500-speaker dataset with a P< 0.001.

This decrement in EER as the number of training speakers increased is visually captured in Fig. 3, which underscores the logarithmic reduction of the error rate with an augmented training set size.

Intelligibility of patients is not an influencing factor in ASV

To explore the relationship between intelligibility of the speakers and ASV, we computed correlation coefficients between EER results (representing speaker verification metric) and WRR values (indicating speech intelligibility) of all the experiments. Figure 4 illustrates the correlation coefficients between error rates and recognition rates of all the experiments. We observed that the correlation coefficients in all cases were very small. Notably, as the number of speakers increased, this correlation diminished even further. Specifically, in the "all-spk-50" experiment-wherein all healthy and pathological speech signals from both children and adults were fused and a random sample of 50 speakers was taken-the correlation coefficient between EER and WRR stood at 0.22 ± 0.30. For larger sample sizes, "all-spk-500" had a coefficient of 0.04 ± 0.09, "all-spk-1500" showed 0.01 ± 0.06, and the largest sample, "all-spk-3000", exhibited an almost non-existent correlation of 0.00 ± 0.04. This data strongly indicates that the intelligibility of a patient's speech does not wield substantial influence over the performance of an ASV system.

Discussion

This study, drawing from in-depth analysis of recordings of both pathological and healthy subjects, offers strong evidence that certain speech pathologies might serve as viable biomarkers in automatic speaker verification (ASV). Intriguingly, certain pathological speech forms demonstrated a heightened vulnerability, shedding light on the potential risks associated with patient re-identification. Using a state-of-the-art deep learning framework for training and evaluation, our research dove deep into these complexities.

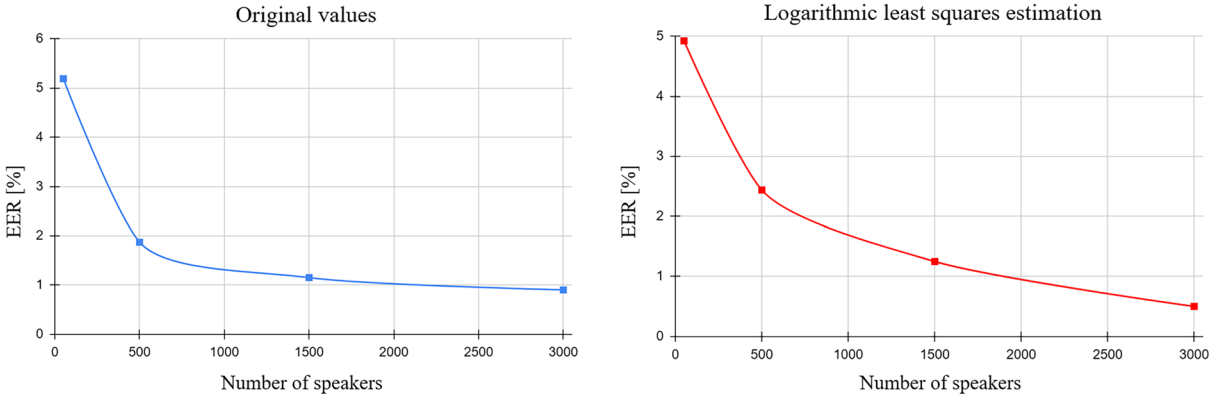

Figure 3. EER results utilizing different training speaker numbers. (**a**) The original values. The EER values are 5.19, 1.87, 1.15, and 0.90 for the cases with n=50, 500, 1500, and 3000 speakers, respectively. (**b**) The resulting curve after logarithmic least squares regression according to $y = 9.1543237903 - 1.0809973418 \cdot \ln x$. The regression coefficient of determination (R^2) equals 0.95. We observe that increasing total training speaker numbers, leads to logarithmic improvement of the ASV performance.

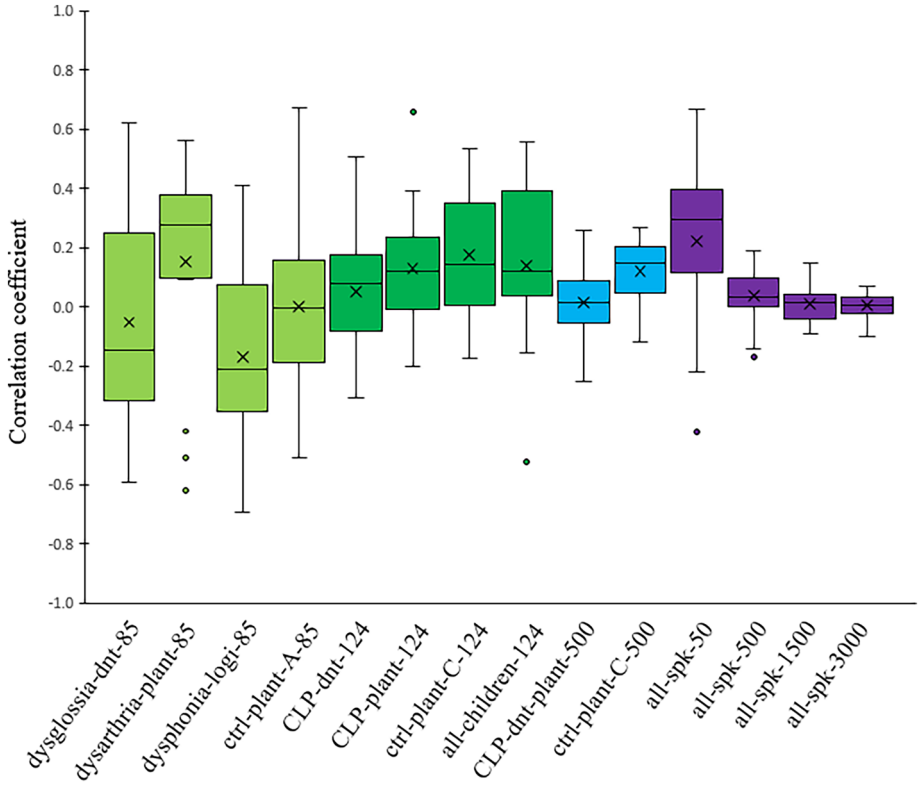

Figure 4. Correlation coefficients between EER values and WRR values for all the experiments. *dysglossia* patients with dysglossia who underwent prior maxillofacial surgery, *dysarthria* patients diagnosed with dysarthria, *dysphonia* patients with voice disorders, *CLP* children with cleft lip and palate, *dnt* recordings from the "dnt Call 4U Comfort" headset[27], *plant* recordings via Plantronics Inc. headset[28], *logi* recordings via Logitech International S.A. headset[29], *ctrl* control group. The labels "-A" and "-C" respectively indicate adult and children subsets. Numbers appended, such as "-85" in "dysglossia-dnt-85", represent the total speaker count for that experiment. "all-spk" designates experiments combining all dataset speech signals from both adults and children, and both pathological and healthy subjects.

To objectively gauge the impact of pathology on ASV, rigorous controls were established to address potential confounders, such as age distribution, recording conditions, microphone types, audio clarity, and speech intelligibility. Analyzing pathological recordings from n=2581 adults and children, the results illustrated a mean EER of $0.89 \pm 0.06\%$. Strikingly, this EER is appreciably lower than that in non-pathological datasets like LibriSpeech[30] or VoxCeleb1 & 2[31,32]. To circumvent biases from random sampling, we derived this result from an average of 20 repeated trials.

Data from children yielded intriguing insights. Pathological children, on average, exhibited higher EER values than their healthy peers. For instance, the "CLP-plant-124" subgroup displayed a 27% surge in EER, under identical recording conditions as the control group. Conversely, adult data showed decreased error rates for those with speech pathologies. This disparity could stem from the ASV model's inclination towards adult speech patterns, coupled with the evolving nature of children's speech influenced by cognitive development.

Our exploration of the relationship between speech pathologies and ASV efficacy yielded further illuminating findings. The integration of diverse pathological voices into the dataset notably enhanced ASV accuracy. For example, the average EER experienced a significant improvement when varied pathologies from the "all-children-124" set were included, performing better than the control group. This suggests that incorporating multiple sources of variability could be pivotal in refining ASV outcomes.

Moreover, the trend of enhanced ASV performance persisted when the training sets were enriched with pathological diversity. For instance, the mixed pathological group's EER was lower than that of the healthy group, emphasizing the potential advantage of pathological diversity in speaker verification.

Delving into the effects of training set size on ASV, we observed that expanding the speaker pool, to include both pathological and healthy voices, consistently boosted ASV accuracy. For example, with the increase in speakers in datasets like "all-spk-500", "all-spk-1500", and "all-spk-3000", there was a consistent drop in EER. Such an incremental improvement with increasing dataset size hints at the potential of large datasets in drastically enhancing ASV efficacy.

Diving deeper into the potential variables that could influence ASV, we probed the intricacies of speech intelligibility. Analyzing the correlation between EER results (indicating ASV performance) and WRR values (indicating speech intelligibility) across experiments, we uncovered intriguing patterns. The consistently minimal

correlation values, especially in larger speaker samples, unequivocally underline that a speaker's intelligibility does not significantly sway ASV system outcomes. This observation challenges the often-presumed importance of speech clarity in ASV systems, suggesting that even if a speaker's utterances are not distinctly clear, it might not substantially hamper the system's verification accuracy. This revelation could have profound implications, especially in scenarios where speech anomalies are prevalent.

Our study stands out due to its novel emphasis on the intersection of speech pathologies and ASV. While a significant portion of recent ASV research has dedicated efforts to improve algorithms and tackle speaker verification challenges by utilizing well-established non-pathological datasets-such as LibriSpeech[30] (EER: 3.85% on the 'test-clean' subset with n=40 test speakers and 3.66% on 'test-other' subset with n=33 test speakers[33]), VoxCeleb 1[31] (EER: 7.80% with n=40 test speakers), and VoxCeleb 2[32] (EER: 3.95% with n=40 test speakers)-there is a conspicuous absence of studies that delve into the relationship between speech pathologies and ASV. In our initial exploration, we identified a substantially low mean EER of $0.89 \pm 0.06\%$ when analyzing pathological speech patterns. While our research introduces a unique dimension to ASV by examining speech pathologies, our results are not directly comparable to those derived from non-pathological conventional datasets because of the inherent differences in the characteristics and challenges posed by pathological speech patterns, recording conditions, testing criteria, text-independent or dependent nature of ASV task, etc. Nonetheless, our study lays the groundwork for a more profound understanding of ASV systems, particularly in contexts permeated by speech anomalies.

Our study had limitations. First, due to the constrained availability of adult subjects, we were unable to harmonize age distributions among individual adult sub-groups, potentially narrowing the generalizability of our findings within adult demographics. To enhance clarity and depth in comparative results, securing additional utterances from both patient and healthy adult populations in future studies is paramount. Second, despite utilizing a robust, large-scale dataset sourced from an extensive array of participants, our pathological corpus[23] was circumscribed to specific speech pathologies and voice disorders, namely dysglossia following maxillofacial surgery, dysarthria, dysphonia, and cleft lip and palate. Subsequent research could potentially broaden this dataset to encompass additional conditions such as aphasia[15]. Furthermore, our pathological corpus[23], though diverse in its recording locations - spanning cities like (i) Erlangen, Bavaria, Germany, (ii) Nuremberg, Bavaria, Germany, (iii) Munich, Bavaria, Germany, (iv) Stuttgart, Baden-Württemberg, Germany, and (v) Siegen, North Rhine-Westphalia, Germany - exclusively features German-language utterances. While we expect that language may not correlate with the susceptibility of pathological speech to re-identification, it remains essential to confirm these findings across multiple languages to validate and generalize our results. Lastly, although we have illuminated the effects of speech pathology across distinct pathology and voice disorder groupings, an important area warranting deeper exploration is the examination at an individual level. In our future direction, this will be a focal area of emphasis.

In conclusion, our findings elucidate the complex relationship between specific speech pathologies and their impact on ASV. We have pinpointed pathologies such as dysphonia and CLP as warranting increased attention due to their amplified re-identification risks. Contrary to prevalent beliefs, our study also reveals that pristine speech clarity is not pivotal for ASV's effective operation. The diversity of datasets plays a crucial role in augmenting ASV performance, a noteworthy insight for future ASV developments. However, as the demand for open-source speech data rises, our study emphasizes the critical need for the development or refinement of anonymization techniques. While research in the domain of anonymization is evolving, as indicated by works like[18,34–36], there remains a pressing need for techniques specifically attuned to pathological speech. It is imperative for the scientific community to strike a harmonious balance between maximizing the utility of data and safeguarding the privacy and rights of individuals.

Methods

Ethics declarations

The study and the methods were performed in accordance with relevant guidelines and regulations and approved by the University Hospital Erlangen's institutional review board (IRB) with application number 3473. Informed consent was obtained from all adult participants as well as from parents or legal guardians of the children.

Pathological speech corpus

Initially, we gathered a total of 216.88 hours of recordings from n=4121 subjects using PEAKS[23], a prominent open-source tool. Given PEAKS' extensive use in scientific circles across German-speaking regions since 2009, its database offers a comprehensive assortment of recordings reflecting a multitude of conditions. To arrive at the finalized dataset, the following steps of intricate analysis were executed:

(i) Recordings missing data points such as WRR, diagnosis, age, microphone, or recording environment were purged from the collection. (ii) Recordings that were noisy or of poor quality were also discarded. (iii) Any data categorized as 'test' or deemed irrelevant by examiners were omitted. (iv) Segments of recordings containing the examiner's voice or those from multiple speakers were excised. (v) Leveraging PEAKS' ability to automatically segment recordings into shorter utterances (ranging from 2 to 10 seconds based on voice activity), speakers that, post these steps, were left with fewer than 8 utterances were excluded. (vi) Finally, recognizing age as a potentially influential variable, the dataset was bifurcated into two major categories: adults and children. This segregation was vital to ensure nuanced analyses given the distinctive characteristics and potential performance deviations associated with these age groups.

In the end, a total of n=3849 participants were included in this study. Table 1 shows an overview and the statistics of the data subsets, i.e., the adults and children. The utilized dataset contained 198.82 hours of recordings from n= 2102 individuals with various pathologies and n=1747 healthy subjects. To ensure our results are

	Total num speakers	Num female speakers	Num male speakers	Total num utterances	Total duration [hours]	Mean ± std dev age [years]	Mean ± std dev WRR [%]
Adults							
dysglossia-dnt	883	245	638	21,338	41.21	60.91 ± 11.95	62.61 ± 15.96
dysarthria-plant	533	258	275	7128	13.37	62.70 ± 15.29	69.11 ± 12.70
dysphonia-logi	86	10	76	900	1.67	59.28 ± 10.67	51.78 ± 15.84
Sum patients	1502	513	989	29,366	56.25	61.40 ± 13.34	63.37 ± 15.78
ctrl-plant-A	85	42	43	891	1.60	23.93 ± 15.62	73.72 ± 15.69
Children							
CLP-dnt	476	216	260	16,964	34.63	9.69 ± 3.98	48.28 ± 17.30
CLP-plant	124	58	66	4120	7.88	9.27 ± 2.58	57.61 ± 13.86
Sum patients	600	264	326	21,084	42.51	9.58 ± 3.71	50.12 ± 17.07
ctrl-plant-C	1662	900	761	54,896	98.46	12.16 ± 3.72	65.87 ± 12.44

Table 1. Dataset statistics used in this study. The table provides details on the total number of speakers, gender distribution, utterance count, total duration in hours, age range, and word recognition rates (WRRs). The corpus is divided into two groups: adults (those aged over 20 years) and children (those aged 20 years or younger). Both groups encompass control subsets ("ctrl-plant") comprising healthy subjects. *dysglossia* patients with dysglossia who had prior maxillofacial surgery before assessment, *dysarthria* patients diagnosed with dysarthria, *dysphonia* patients with voice disorders, *CLP* children diagnosed with cleft lip and palate, *dnt* recordings using the "dnt Call 4U Comfort" headset[27], *plant* recordings using a specific Plantronics Inc. headset[28], *logi* recordings using a specific Logitech International S.A. headset[29], *ctrl* control group. The suffix "-A" denotes the adult subset, whereas "-C" pertains to the children subset. Age and WRR metrics are expressed as mean ± standard deviation.

reliable, we carefully sorted these recordings based on pathology types and recording settings. The utterances were recorded at 16kHz sampling frequency and 16 bit resolution[23]. All the subjects were native German speakers, using different dialects including the standard German ("Hochdeutsch") as well as local dialects.

Adults

Subjects above the age of 20 were included in the adults subset of our dataset. n=1,502 patients read "Der Nordwind und die Sonne", the German version of the text "The North Wind and the Sun", a fable from Aesop. It is a phonetically rich text with 108 words, of which 71 are unique[23]. Our adult patient cohort had an age range of 21 to 94 years (mean 61.40 ± 13.34 and median 62.49). Figure 5a shows the age histogram of the three patient groups of adults used in this study ("dysglossia-dnt", "dysarthria-plant", and "dysphonia-logi").

"dysglossia-dnt" represents the group of patients who had dysglossia, underwent a maxillofacial surgery before the pathology assessment, and all were recorded using the "dnt Call 4U Comfort" headset[27]. Out of all the available utterances, we selected those that were recorded using the same microphone. "dysarthria-plant"

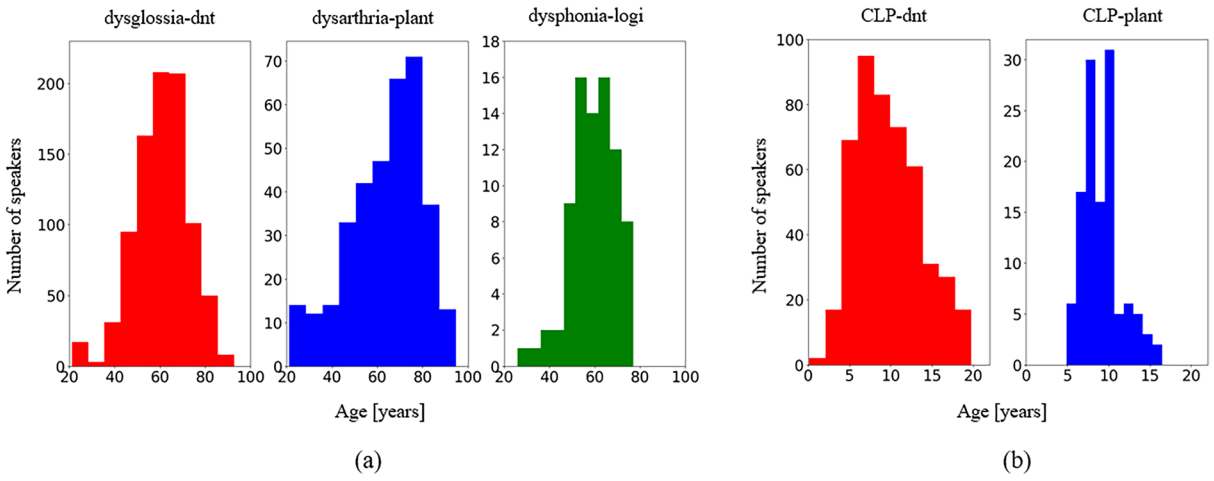

Figure 5. Age histograms of the patient groups. (**a**) The adults; (**b**) The children group. *dysglossia* patients with dysglossia who underwent prior maxillofacial surgery. *dysarthria* patients diagnosed with dysarthria, *dysphonia* patients with voice disorders, *CLP* children with cleft lip and palate, *dnt* recordings from the "dnt Call 4U Comfort" headset[27], *plant* recordings via Plantronics Inc. headset[28], *logi* recordings via Logitech International S.A. headset[29].

Experiment name	Total num speakers	Subset	Pathology	Microphone
dysglossia-dnt-85	85	Adults	Dysglossia	dnt Call 4U
dysarthria-plant-85	85	Adults	Dysarthria	Plantronics
dysphonia-logi-85	85	Adults	Dysphonia	Logitech
ctrl-plant-A-85	85	Adults	None (healthy)	Plantronics
CLP-dnt-124	124	Children	CLP	dnt Call 4U
CLP-plant-124	124	Children	CLP	Plantronics
ctrl-plant-C-124	124	Children	None (healthy)	Plantronics
all-children-124	124	Children	Mixture of CLP & healthy	dnt Call 4U & Plantronics
CLP-dnt-plant-500	500	Children	CLP	dnt Call 4U
ctrl-plant-C-500	500	Children	None (healthy)	Plantronics
all-spk-50	50	Adults & children	Mixture of all & healthy	Mixture of all
all-spk-500	500	Adults & children	Mixture of all & healthy	Mixture of all
all-spk-1500	1500	Adults & children	Mixture of all & healthy	Mixture of all
all-spk-3000	3000	Adults & children	Mixture of all & healthy	Mixture of all

Table 2. Overview of the experiments performed in this study. *dysglossia* patients with dysglossia who underwent prior maxillofacial surgery, *dysarthria* patients diagnosed with dysarthria, *dysphonia* patients with voice disorders, *CLP* children with cleft lip and palate, *dnt* recordings from the "dnt Call 4U Comfort" headset[27], *plant* recordings via Plantronics Inc. headset[28], *logi* recordings via Logitech International S.A. headset[29], *ctrl* control group. The labels "-A" and "-C" respectively indicate adult and children subsets. Numbers appended, such as "-85" in "dysglossia-dnt-85", represent the total speaker count for that experiment. "all-spk" designates experiments combining all dataset speech signals from both adults and children, and both pathological and healthy subjects.

is a group of patients who had dysarthria and underwent speech therapy and all were recorded using a specific headset from Plantronics Inc.[28]. "dysphonia-logi" represents the patients who had voice disorders and all were recorded using a specific headset from Logitech International S.A.[29]. Finally, as a control group ("ctrl-plant-A"), n=85 healthy individuals were asked to undergo the test using the same Plantronics headset[28].

Children

Six hundred children with an age range of $2-20$ years old (mean 9.58 ± 3.71 and median 9.12) were included in the study. The test, namely the PLAKSS test, consisted of slides that showed pictograms of the words to be named. In total, the test contained 97 words which included all German phonemes in different positions. Due to the fact that some children tended to explain the pictograms with multiple words, and some additional words were uttered in between the target words, the recordings were automatically segmented at pauses that were longer than 1s[23]. Figure 5b illustrates the age histogram of the two patient groups of children used in this study ("CLP-dnt" and "CLP-plant").

"CLP-dnt" represents children with cleft lip and palate (CLP), which is the most common malformation of the head with incomplete closure of the cranial vocal tract[27,37–39], which all were recorded using the same "dnt Call 4U Comfort" headset[27] as for the adults. Finally, as a control group ("ctrl-plant-C"), n=1,662 healthy children were asked to undergo the test with similar recording conditions as in "ctrl-plant-A".

Experimental design

Table 2 shows an overview of the different experiments performed in this study.

Analysis of impact of pathology on ASV performance

Initially, the study aimed to analyze the performance of automatic speaker verification (ASV) systems on recordings from individuals with various speech pathologies. For each category of adults, recordings were sourced from 85 predetermined speakers. As reflected in Table 1, a precise age match for adults was challenging due to the limited recordings available. Nonetheless, 20% of the speakers were assigned to the test set and 80% to the training set. This selection and allocation process was iterated 20 times. For the children's group, given the limited population size of the "CLP-plant" subgroup as seen in Table 1, recordings from $n = 124$ speakers were chosen, aiming for an average age close to 9.30 ± 2.60. These speakers were similarly divided, with 20% for testing and 80% for training, and this procedure was repeated 20 times.

Effect of pathological diversity

The study further investigated the influence of pathology diversity on speaker verification performance. Consistent with data in Fig. 2, the same number of speakers for both training and testing was maintained, with a focus on closely matching age distribution. By pooling all patient data, the study contrasted the results against a control group. As indicated in Table 1, for children, both age and size consistency were achievable due to the extensive recordings from healthy subjects. Following the established protocol, 20 iterations were conducted where $n = 400$ speakers with a mean age of $10.29 \pm 0.13\%$ and a mean total duration of $26.55 \pm 0.58\%$ were

selected for training. Meanwhile, 100 speakers with a mean age of $10.05 \pm 0.48\%$ and a mean total duration of $6.80 \pm 0.30\%$ were designated for testing from the combined "CLP-dnt" and "CLP-plant" patient groups. Concurrently, 400 speakers with a mean age of $11.72 \pm 0.10\%$ and a mean total duration of $24.08 \pm 0.55\%$ for training and $n = 100$ speakers with a mean age of $11.70 \pm 0.33\%$ and a mean total duration of $6.03 \pm 0.32\%$ for testing were chosen from the "ctrl-plant-C" group.

Training size's influence
This section explored the effect of training set size on ASV system performance. Using recordings from different patient groups alongside a control set, the selection was determined by age and recording duration. To specifically assess training size impact, all $n = 3849$ available pathological and healthy speakers were amalgamated. Different quantities of speakers were randomly chosen for the routine training and evaluation steps: $n = 50$, 500, 1500, and 3000 speakers. For each group, 20% was allocated to the test set and 80% to the training set. Each sampling and evaluation cycle was reiterated 20 times to consider random variations.

Intelligibility's effect
The final phase was a correlation analysis, aiming to discern the relationship between speaker clarity (measured by intelligibility metrics) and ASV system performance metrics. This correlation explored the connection between EER results and WRR values throughout all experimental stages, offering insights into pathological speech nuances within speaker verification systems.

DL-based ASV system

Although DL-based methods, generally, outperform the classical speaker recognition methods, for instance, the i-vector approach[40–42], in the context of text-independent speaker verification (TISV), the i-vector framework and its variants are still the state-of-the-art in some of the tasks[43–46]. However, i-vector systems showed performance degradation when short utterances are met in enrollment/evaluation phase[44]. Given that the children subset of our corpus contains a large amount of utterances with short lengths (less than 4 s), due to the nature of the PLAKSS test it makes sense for us to select a generalized TISV model, which can address our problem better. According to the results reported in[44,47,48], end-to-end DL systems achieved better performance compared to the baseline i-vector system[41], especially for short utterances. A major drawback of these systems is the time and cost required for training. Because of the nature of this study, we aimed at performing a considerable number of different experiments. Therefore, having a state-of-the-art end-to-end TISV model, which requires less training time is crucial. Thus, we chose to utilize the Generalized End-to-End (GE2E) TISV model proposed by Wan et al.[49], which enabled us to process a large number of utterances at once and greatly decreased the total training and convergence time[33]. The final embedding vector (d-vector) $\mathbf{e}_{ji}$ was the L_2 normalization of the network output and represents the embedding vector of the jth speaker's ith utterance. The centroid of the embedding vectors from the jth speaker $[\mathbf{e}_{j1}, \ldots, \mathbf{e}_{jM}]$ $\mathbf{c}_j$ was defined as the arithmetic mean of the embedding vectors of the jth speaker.

The similarity matrix $\mathbf{S}_{ji,k}$ was defined as the scaled cosine similarities between each embedding vector $\mathbf{e}_{ji}$ to all centroids $\mathbf{c}_k$ ($1 \le j, k \le N$, and $1 \le i \le M$). Furthermore, removing $\mathbf{e}_{ji}$ when computing the centroid of the true speaker made training stable and helps avoid trivial solutions[49]. Thus, the similarity matrix could be written as following:

$$\mathbf{S}_{ji,k} = \begin{cases} w \cdot \cos(\mathbf{e}_{ji}, \mathbf{c}_j^{(-i)}) + b & \text{if } k = j \\ w \cdot \cos(\mathbf{e}_{ji}, \mathbf{c}_k) + b & \text{otherwise,} \end{cases} \tag{1}$$

with w and b being the trainable weights and biases. As we can see, unlike most of the end-to-end methods, rather than a scalar value, GE2E builds a similarity matrix that defines the similarities between each $\mathbf{e}_{ji}$ and all centroids $\mathbf{c}_k$.

We put a SoftMax on $\mathbf{S}_{ji,k}$ for $k = 1, \ldots, N$ that makes the output equal to one if $k = j$, otherwise makes the output equal to zero. Thus, the loss on each embedding vector $\mathbf{e}_{ji}$ could be defined as:

$$L(\mathbf{e}_{ji}) = -\mathbf{S}_{ji,j} + \log \sum_{k=1}^{N} \exp(\mathbf{S}_{ji,k}). \tag{2}$$

Finally, the GE2E loss L_G is the mean of all losses over the similarity matrix ($1 \le j \le N$, and $1 \le i \le M$):

$$L_G = \frac{1}{M \cdot N} \sum_{j,i} L(\mathbf{e}_{ji}). \tag{3}$$

Training steps

By specifying a set of clear training and evaluation steps for all the experiments, we aimed at standardizing our experiments and preventing influences of non-pathology factors. We followed a similar data pre-processing scheme as in[33,49,50] and pruned the intervals with sound pressures below 30 db. Afterward, we performed voice activity detection[51] to remove the silent parts of the utterances, with a window length of 30 ms, a maximum silence length of 6 ms, and a moving average window of the length 8 ms. Removing silent parts, we ended up with partial utterances of each utterance, where we merely chose the partial utterances which have a minimum length of 1825 ms for training, due to the fact that our dataset contained utterances with a 16 kHz sampling rate. Our final feature representations were 40-dimensional log-Mel-filterbank energies, where we used a window length

```
for all training batches do
    – randomly choose an integer L within [140, 180];
    for all train speakers do
        – randomly choose N speakers;
        for all N speakers do
            – initialize empty set S;
            for all utterances do
                – normalize the volume;
                – perform VAD with max_silence_length = 6 ms and window_length= 30 ms;
                – prune the intervals with sound pressures below 30 db;
                for all resulting partial utterances do
                    if partial utterance's length >180 frames then
                        – add partial utterance to S;
            – randomly select M partial utterances from S;
            for all selected partial utterances do
                – perform STFT on the partial utterance;
                – take magnitude squared of result;
                – transform to the Mel scale;
                – take the logarithm;
                – randomly segment an interval with L frames;
```

Algorithm 1. Training data preparation steps.

of 25 ms with steps of 10 ms and, i.e., a short time Fourier transform (STFT) of size of 512. To prepare training data batches, similar to[49], we selected N different speakers and fetched M different utterances for every selected speaker to create a training batch. Furthermore, while we could have batches with different partial utterance lengths, due to the fact that all the partial utterances of each training batch should have the same length[49], we randomly segmented all the partial utterances of each training batch to have the same length. All the procedures to pre-process raw input waveforms and prepare training batches are explained in Algorithm 1.

Our network architecture, which is shown in Fig. 6, consisted of 3 long short-term memory (LSTM) layers[52] with 768 hidden nodes followed by a linear projection layer in order to get to the 256-dimensional embedding vectors[53]. The L_2 norm of gradient was clipped at 3[54]. In order to prevent coincidental training cases, the Xavier normal initialization[55] was applied to the network weights and the biases were initialized with zeros for all the experiments. The Adam[56] optimizer was selected to optimize the model. Depending on each individual experiment and more specifically, its training set, we chose a different learning rate per experiment from 10^{-5} to 10^{-4}, in a way that the network converges the best. For all of the experiments, during training, we selected $N = 16$ speakers and $M = 4$ partial utterances per speaker. Moreover, no pre-trained model was used during training of each experiment, and we always started training from scratch with the same initialization.

Evaluation method

For the evaluation of the trained networks, we followed the same data pre-processing steps as for training, with the only difference that, during evaluation, we concatenated all the partial utterances corresponding to each utterance before feeding them to the network. Then, as proposed by Wan et al.[49], we applied a sliding window of a fixed size (160 frames) with 50% overlap to the concatenated utterances and performed an element-wise averaging on the d-vectors to get the final d-vector representation of the test utterance. Furthermore, Tayebi Arasteh[33] showed that the choice of the parameter M for evaluation is an influencing factor in the resulting prediction, i.e., the more enrollment utterances, usually, the better prediction for test utterances. Therefore, we decided to report the results for $M = 2$, where we have only one enrollment utterance (during the calculation of centroid of the true speaker, we excluded the utterance itself as proposed by Wan et al.[49]), as we did not see large deviations for other choices of M in results which cannot be reported here for brevity. The results for $M = 4$ are

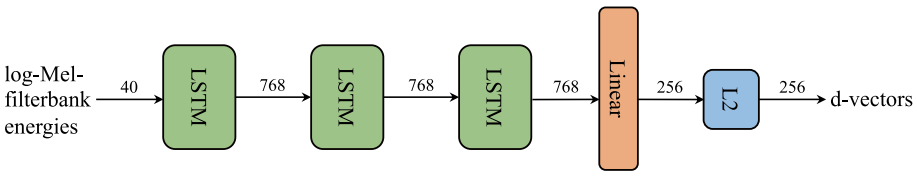

Figure 6. The architecture of the utilized text-independent speaker verification model. The inputs of the network are 40-dimensional log-Mel-filterbank energies, which are the results of performing data pre-processing steps on raw utterances. The numbers above each arrow represent the feature dimensions at each step. The final 256-dimensional d-vectors are the L_2 normalization of the network outputs.

```
for all enrollment and evaluation speakers do
    for all utterances do
        – initialize empty set A;
        – normalize the volume;
        – perform VAD with max_silence_length = 6 ms and window_length = 30 ms;
        – prune the intervals with sound pressures below 30 db;
        for all resulting partial utterances do
            if partial utterance's length > 180 frames then
                – add the partial utterance to A;
        – concatenate the elements of A;
        – perform STFT on the concatenated utterance;
        – take the magnitude squared of the result;
        – transform to the Mel scale;
        – take the logarithm;
        – set t = 0;
        – initialize empty set D;
        while t + 160 < length of the utterance do
            – select the interval within [t, t + 160] frames of the utterance;
            – feed the selected utterance to the trained network to obtain the corresponding d-vector;
            – L2-normalize the d-vector;
            – add the normalized d-vector to D;
            – t = t + 80;
        – perform element-wise average of elements of D to obtain the final utterance d-vector;
```

Algorithm 2. Enrollment and evaluation data preparation followed by d-vector creation steps.

reported in the supplementary information (see Table S1). For each experiment, we chose the batch size N to be equal to the total number of the test speakers during evaluation. To prevent the effect of random sampling in choosing recordings of training and testing for different experiments, we repeated each experiment 20 times and calculated the statistics accordingly. All the steps to pre-process raw input waveforms for enrollment and evaluation as well as the steps for preparing final d-vectors are stated in Algorithm 2.

Quantitative analysis metric

As our main quantitative evaluation metric, we chose EER, which is used to predetermine the threshold values for its false acceptance rate (FAR) and its false rejection rate (FRR)[57,58]. It looks for a threshold for similarity scores where the proportion of genuine utterances which are classified as imposter, i.e., the FRR is equal to the proportion of imposters classified as genuine, i.e., the FAR[33]. The similarity metric, which we use here, is the cosine distance score, which is the normalized dot product of the speaker model and the test d-vector:

$$\cos(\mathbf{e}_{ji}, \mathbf{c}_k) = \frac{\mathbf{e}_{ji} \cdot \mathbf{c}_k}{\|\mathbf{e}_{ji}\| \cdot \|\mathbf{c}_k\|}. \tag{4}$$

The higher the similarity score between $\mathbf{e}_{ji}$ and $\mathbf{c}_k$ is, the more similar they are. We report the EER values in percent throughout this paper.

Statistical analysis

Descriptive statistics are reported as median and range, or mean ± standard deviation, as appropriate. Normality was tested using Shapiro-Wilk test[59]. A two-tailed unpaired t-test was used to compare two groups of EER data with Gaussian distributions. A $P \leqslant 0.05$ was considered statistically significant.

Hardware

The hardware used in our experiments were Intel CPUs with 18 and 32 cores and 32 GB RAM and Nvidia GPUs of GeForce GTX 1080 Ti, V100, RTX 6000, Quadro 5000, and Quadro 6000 with 11 GB, 16 GB, 24 GB, 32 GB, and 32 GB memories, respectively.

Data availability

The speech dataset used in this study is not publicly available as it is internal data of patients of the University Hospital Erlangen. A reasonable request to the corresponding author is required for accessing the data on-site at the University Hospital Erlangen in Erlangen, Bavaria, Germany.

Code availability

The full source code including training and evaluation of the recurrent neural networks, data pre-processing and feature extraction steps, and analysis of the results are publicly available at https://github.com/tayebiarasteh/pathology_ASV. All the code is developed in Python 3.9. The PyTorch 1.13 framework is used for deep learning.

Received: 2 December 2022; Accepted: 17 November 2023
Published online: 22 November 2023

References

1. Rios-UrregoEmail, C., Vásquez-Correa, J., Orozco-Arroyave, J. & Nöth, E. Is there any additional information in a neural network trained for pathological speech classification? In *Proc. 24th International Conference on Text, Speech, and Dialogue, Olomouc, Czech Republic*, 435–447, https://doi.org/10.1007/978-3-030-83527-9_37 (Springer Nature, 2021).
2. Sztahó, D., Szaszák, G. & Beke, A. Learning methods in speaker recognition: A review. *Period. Polytech. Electr. Eng. Comput. Sci.* **65**, 310–328. https://doi.org/10.3311/PPee.17024 (2021).
3. Moro-Velazquez, L., Villalba, J. & Dehak, N. Using x-vectors to automatically detect parkinson's disease from speech. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1155–1159 (2020).
4. Pappagari, R., Cho, J., Moro-Velázquez, L. & Dehak, N. Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity. In *Proc. INTERSPEECH 2020*, 2177–2181, https://doi.org/10.21437/Interspeech.2020-2587 (2020).
5. Moro-Velazquez, L. *et al.* Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's disease. *Appl. Soft Comput.* **62**, 649–666. https://doi.org/10.1016/j.asoc.2017.11.001 (2018).
6. Tayebi Arasteh, S. *et al.* Federated Learning for Secure Development of AI Models for Parkinson's Disease Detection Using Speech from Different Languages. In *Proc. INTERSPEECH 2023*, 5003–5007, https://doi.org/10.21437/Interspeech.2023-2108 (2023).
7. Jamal, N., Shanta, S., Mahmud, F. & Sha'abani, M. Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review. *AIP Conf. Proc.* **1883**, 020028. https://doi.org/10.1063/1.5002046 (2017).
8. Demir, K. C. *et al.* Pocap corpus: A multimodal dataset for smart operating room speech assistant using interventional radiology workflow analysis. In *Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic, September 6-9, 2022, Proceedings*, 464-475, https://doi.org/10.1007/978-3-031-16270-1_38 (Springer-Verlag, 2022).
9. Yang, S. H. & Chung, M. Improving dysarthric speech intelligibility using cycle-consistent adversarial training. Preprint at arXiv:2001.04260https://doi.org/10.48550/arXiv.2001.04260 (2020).
10. Maier, A. *Speech of Children with Cleft Lip and Palate: Automatic Assessment* (Logos-Verlag, 2009).
11. Vásquez-Correa, J. C. *et al.* Multimodal assessment of Parkinson's disease: A deep learning approach. *IEEE J. Biomed. Health Inform.* **23**, 1618–1630. https://doi.org/10.1109/JBHI.2018.2866873 (2019).
12. Rios-Urrego, C. D. *et al.* Automatic pronunciation assessment of non-native English based on phonological analysis. In *Text, Speech, and Dialogue* (eds Ekštein, K. *et al.*) 339–348 (Springer Nature Switzerland, 2023).
13. Pérez-Toro, P. A. *et al.* Transferring quantified emotion knowledge for the detection of depression in alzheimer's disease using forestnets. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5, https://doi.org/10.1109/ICASSP49357.2023.10095219 (2023).
14. García, M. A. & Rosset, A. L. Deep neural network for automatic assessment of dysphonia. Preprint at arXiv:2202.12957https://doi.org/10.48550/arXiv.2202.12957 (2022).
15. Kohlschein, C., Schmitt, M., Schüller, B., Jeschke, S. & Werner, C. J. 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)*, 1–6, https://doi.org/10.1109/HealthCom.2017.8210766 (2017).
16. Bhat, C. & Strik, H. Automatic assessment of sentence-level dysarthria intelligibility using BLSTM. *IEEE J. Sel. Topics Signal Process.* **14**, 322–330. https://doi.org/10.1109/JSTSP.2020.2967652 (2020).
17. Gargot, T. *et al.* Automatic assessment of motor impairments in autism spectrum disorders: A systematic review. *Cogn. Comput.* **14**, 624–659. https://doi.org/10.1007/s12559-021-09940-8 (2022).
18. Tomashenko, N. *et al.* The voiceprivacy 2020 challenge: Results and findings. *Comput. Speech Lang.* **74**, 101362. https://doi.org/10.1016/j.csl.2022.101362 (2022).
19. Strimbu, K. & Tavel, J. What are biomarkers?. *Curr. Opin. HIV AIDS* **5**, 463–6. https://doi.org/10.1097/COH.0b013e32833ed177 (2010).
20. Califf, R. M. Biomarker definitions and their applications. *Exp. Biol. Med.* **243**, 213–221. https://doi.org/10.1177/1535370217750088 (2018).
21. Marmar, C. R. *et al.* Speech-based markers for posttraumatic stress disorder in us veterans. *Depress. Anxiety* **36**, 607–616. https://doi.org/10.1002/da.22890 (2019).
22. Ramanarayanan, V., Lammert, A. C., Rowe, H. P., Quatieri, T. F. & Green, J. R. Speech as a biomarker: Opportunities, interpretability, and challenges. *Perspect. ASHA Spec. Interest Groups* **7**, 276–283. https://doi.org/10.1044/2021_PERSP-21-00174 (2022).
23. Maier, A. *et al.* Peaks - A system for the automatic evaluation of voice and speech disorders. *Speech Commun.* **51**, 425–437. https://doi.org/10.1016/j.specom.2009.01.004 (2009).
24. Kinnunen, T. & Li, H. An overview of text-independent speaker recognition: From features to supervectors. *Speech Commun.* **52**, 12–40. https://doi.org/10.1016/j.specom.2009.08.009 (2010).
25. Bimbot, F. *et al.* A tutorial on text-independent speaker verification. *EURASIP J. Adv. Signal Process.*https://doi.org/10.1155/S1110865704310024 *(2004).*
26. Kitzing, P., Maier, A. & Åhlander, V. L. Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders. *Logop. Phoniatr. Vocol.* **34**, 91–96. https://doi.org/10.1080/14015430802657216 (2009).
27. Maier, A., Noeth, E., Batliner, A., Nkenke, E. & Schuster, M. Fully automatic assessment of speech of children with cleft lip and palate. *Informatica (Slovenia)* **30**, 477–482 (2006).
28. Plantronics inc., Santa cruz, CA, USA. https://www.poly.com/.
29. Logitech international s.a., Lausanne, Switzerland. https://www.logitech.com/.
30. Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210 (2015).
31. Nagrani, A., Chung, J. S. & Zisserman, A. VoxCeleb: A large-scale speaker identification dataset. *Proc. Interspeech 2017*https://doi.org/10.21437/Interspeech.2017-950 *(2017).*
32. Chung, J. S., Nagrani, A. & Zisserman, A. VoxCeleb2: Deep speaker recognition. *Proc Interspeech 2018*https://doi.org/10.21437/Interspeech.2018-1929 *(2018).*
33. Tayebi Arasteh, S. An empirical study on text-independent speaker verification based on the ge2e method. Preprint at arXiv:2011.04896https://doi.org/10.48550/arXiv.2011.04896 (2020).
34. Perero-Codosero, J. M., Espinoza-Cuadros, F. M. & Hernández-Gómez, L. A. X-vector anonymization using autoencoders and adversarial training for preserving speech privacy. *Comput. Speech Lang.* **74**, 101351. https://doi.org/10.1016/j.csl.2022.101351 (2022).
35. Yoo, I.-C. *et al.* Speaker anonymization for personal information protection using voice conversion techniques. *IEEE Access* **8**, 198637–198645. https://doi.org/10.1109/ACCESS.2020.3035416 (2020).
36. Srivastava, B. M. L. *et al.* Design choices for x-vector based speaker anonymization. Preprint at arXiv:2005.08601https://doi.org/10.48550/arXiv.2005.08601 (2020).

37. Wantia, N. & Rettinger, G. The current understanding of cleft lip malformations. *Facial Plast. Surg. FPS* **18**, 147–53. https://doi.org/10.1055/s-2002-33061 (2002).
38. Millard, T. & Richman, L. Different cleft conditions, facial appearance, and speech: Relationship to psychological variables. *The Cleft Palate-Craniofacial J. .Off. Publ. Am. Cleft Palate-Craniofacial Assoc.*https://doi.org/10.1597/1545-1569_2001_038_0068_dccfaa_2.0.co_2 *(2001).*
39. Harding, A. & Grunwell, P. Characteristics of cleft palate speech. *Int. J. Lang. Commun. Disord.* **31**, 331–357. https://doi.org/10.3109/13682829609031326 (1996).
40. Dehak, N. *et al.* Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. *Proc. Interspeech 2009*https://doi.org/10.21437/Interspeech.2009-385 *(2009).*
41. Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P. & Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* **19**, 788–798. https://doi.org/10.1109/TASL.2010.2064307 (2011).
42. Dehak, N. *et al.* Support vector machines and joint factor analysis for speaker verification. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4237–4240, https://doi.org/10.1109/ICASSP.2009.4960564 (2009).
43. Lei, Y., Scheffer, N., Ferrer, L. & McLaren, M. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1695–1699, https://doi.org/10.1109/ICASSP.2014.6853887 (2014).
44. Zhang, C. & Koishida, K. End-to-end text-independent speaker verification with triplet loss on short utterances. *Proc. Interspeech 2017*https://doi.org/10.21437/Interspeech.2017-1608 *(2017).*
45. Nist speaker recognition evaluation 2012. http://www.nist.gov/itl/iad/mig/sre12.cfm (2012).
46. Nist speaker recognition evaluation 2016. https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016 (2016).
47. Snyder, D. *et al.* Deep neural network-based speaker embeddings for end-to-end speaker verification. *2016 IEEE Spoken Lang. Technol. Workshop (SLT)*https://doi.org/10.1109/SLT.2016.7846260 *(2016).*
48. Bredin, H. Tristounet: Triplet loss for speaker turn embedding. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5430–5434, https://doi.org/10.1109/ICASSP.2017.7953194 (2017).
49. Wan, L., Wang, Q., Papir, A. & Moreno, I. L. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4879–4883, https://doi.org/10.1109/ICASSP.2018.8462665 (2018).
50. Prabhavalkar, R., Alvarez, R., Parada, C., Nakkiran, P. & Sainath, T. N. Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4704–4708 (2015).
51. Ramirez, J., Gorriz, J. M. & Segura, J. C. Voice activity detection, fundamentals and speech recognition system robustness. In *Robust Speech* Vol. 1 (eds Grimm, M. & Kroschel, K.) (IntechOpen, 2007). https://doi.org/10.5772/4740.
52. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 (1997).
53. Sak, H., Senior, A. & Beaufays, F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *Proc. Interspeech 2014*https://doi.org/10.21437/Interspeech.2014-80 *(2014).*
54. Pascanu, R., Mikolov, T. & Bengio, Y. On the difficulty of training recurrent neural networks. *Proc. 30th Int. Conf. Mach. Learn. PMLR* **28**, 1310–1318 (2013).
55. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. *Proc. Thirteenth Int. Conf. Artif. Intell. Stat. PMLR* **9**, 249–256 (2010).
56. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations (ICLR)*, 4704–4708 (2015).
57. van Leeuwen, D. A. & Brümmer, N. *An Introduction to Application-Independent Evaluation of Speaker Recognition Systems, 330–353* (Springer, 2007).
58. Hansen, J. H. L. & Hasan, T. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Process. Mag.* **32**, 74–99. https://doi.org/10.1109/MSP.2015.2462851 (2015).
59. Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika* **52**, 591–611. https://doi.org/10.2307/2333709 (1965).

Author contributions

S.T.A. cleaned, organized, and pre-processed the data, developed the software, conducted the experiments, performed the statistical analysis, and wrote the manuscript. T.W. prepared the data and contributed to the editing. M.S. guided the data collection, counseled on clinical relevance, and contributed to the editing. E.N. guided the study design and contributed to the editing. A.M. supported the conception of the study and the experiments, corrected, and edited the manuscript. S.H.Y. designed the study and greatly guided the writing, corrected, and edited the manuscript.

Funding

We acknowledge financial support by Deutsche Forschungsgemeinschaft and Friedrich-Alexander-Universität Erlangen-Nürnberg within the funding programme "Open Access Publication Funding". Open Access funding enabled and organized by Projekt DEAL. This study was funded by Friedrich-Alexander-University Erlangen-Nuremberg, Medical Valley e.V., and Siemens Healthineers AG within the framework of d.hip campus.

Competing Interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-47711-7.

Correspondence and requests for materials should be addressed to S.T.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.