To appear in Proc. EAAI 2026

Speaker Anonymization for Children's Oral Reading Assessment

Sandipan Dhar, Srikanth Raj Chetupalli, and Preeti Rao

Indian Institute of Technology Bombay, India
sandipandhartsk03@gmail.com, srikanthrajch@iitb.ac.in, prao@ee.iitb.ac.in

Abstract

Speaker anonymization aims to modify the speech signal in order to protect the identity of a speaker while preserving the linguistic content. Despite the increasing use of children's voices in educational applications, such as oral reading fluency (ORF) assessment, there is little work on the anonymization aspects. In this work, we investigate the effectiveness of available speaker anonymization methods drawing from traditional speech-production based approaches and a neural codec based method. We investigate the trade-off between privacy protection, measured as the degree of anonymity, and utility preservation, which in the current context of ORF assessment, includes the segmental and suprasegmental features of children's read speech utterances. We report objective and subjective evaluations using two child-speaker datasets: MPS and SpeechOcean. Our objective evaluation results indicate that the speech-production based method of vocal tract length normalization coupled with pitch-transposition achieves the best balance between privacy and utility. Subjective listening results indicate that naturalness is achievable across methods while the neural method fails to preserve age characteristics, which are more easily controlled by the speech-production driven methods.

Introduction

The rapid advancement of digital technologies has made privacy concerns dominant, particularly in the context of interactions involving children (Kulkarni et al. 2025). The increasing accessibility of artificial speech generation tools has further intensified these concerns, as it is now easier than ever to misuse an individual's personal traits. Children, in particular, are more vulnerable to such risks due to their limited awareness of digital threats. Among the many privacy-related concerns, voice privacy emerges as a significant issue, as children's speech is increasingly collected through smart devices, educational apps, and other voice-enabled technologies (Evanini and Wang 2013).

A prominent way to address the voice privacy concern is speaker anonymization (Yoo et al. 2020), which aims to modify the speaker's identity while preserving the original utility of the signal, which in the case of a speech recognition problem is the speech content. Prevalent approaches first disentangle the semantic and speaker related attributes, and then modify the speaker attributes to achieve anonymization, often choosing the output speaker from a pseudo speaker set (Meyer, Lux, and Vu 2024; Meyer et al. 2023). However, these systems are trained and evaluated mostly on the adult speech and their performance on children's speech applications is less studied. This highlights the importance of research on children's voice privacy, especially in the era of AI where educational platforms increasingly rely on student computer interactions.

Recently, Kulkarni et al. (2025) investigated the effectiveness of various anonymization methods ranging from traditional source-filter model-based techniques to deep learning based approaches developed in the context of the Voice Privacy Challenge (VPC) (Tomashenko et al. 2022, 2024) for anonymizing child speakers. However, this study overlooked critical factors, such as the effect of speech formant shifting or pitch transposition on privacy and utility. Arasteh et al. (2024) applied speaker anonymization in the context of a speech pathology classification task involving both children and adult speakers. The authors studied pitch transformation with a HiFi-GAN based vocoder (Kong, Kim, and Bae 2020) and compared it with the McAdams Coefficient method (Patino et al. 2021). They reported that signal processing based anonymization offers a better balance between privacy and utility for pathological speech classification. A key requirement in child speaker anonymization is to achieve anonymization while preserving child like vocal characteristics, rather than generating the overly adult like traits often produced by deep learning based anonymization models. This is desirable, for example, in assessment scenarios where the data is accessible to machine as well as human listeners such as teachers when required.

This paper focuses on voice privacy in the context of measuring the Oral Reading Fluency (ORF) of children. Hence, in this work the utility corresponds to the spoken content (i.e. pronunciation of words) as well as the prosodic attributes of the original signal. The preservation of age related vocal characteristics in children is considered as an additional utility requirement. Given this, we study signal processing methods that effect speaker voice change by controlled spectral envelope warping or by pitch shifting. The modifications can be implemented in a frame-based analysis-synthesis framework such as the WORLD vocoder

(Morise, Yokomori, and Ozawa 2016). Spectral envelope warping can be achieved by Vocal Tract Length Normalization (VTLN) methods or by the McAdams coefficient method (Patino et al. 2021), both of which essentially lead to shifting the formant locations in the speech frame by a specified amount. The use of the analysis-by-synthesis framework further facilitates modification of pitch. As a representative of recent deep learning methods, we also investigate Neural Audio Codec Language Modeling (NACLM) (Panariello et al. 2023) approach. We analyze how each of the considered methods affects privacy and utility in the anonymized speech. We report objective and subjective experiments using two publicly available child speaker datasets: Maharashtra Primary Schools (MPS) dataset (Gothi et al. 2024) and the SpeechOcean dataset (Zhang et al. 2021). Given the ubiquity of background noise in our application and the known sensitivity to noise of anonymization methods, we study the performance of our methods on noisy samples from the MPS dataset alongside their enhanced versions.

Speaker Anonymization in ORF Assessment

The ORF assessment task involves processing the audio recording of grade-appropriate text read aloud by a child speaker for the established rubrics of accuracy, pace and expressiveness (Bailly et al. 2022). In a typical system, an automatic speech recognition (ASR) module is used to obtain the text hypothesis for the uttered words, which is then compared with the reference to estimate the number of correctly read words in the given time. A prosody analysis for phrase boundaries and prominence is used to obtain an expressiveness score. The scores are made available to the teacher together with the audio recording. Figure 1 illustrates this framework, considering three scenarios: without anonymization, with strong anonymization, and with balanced anonymization.

In case 1, where speaker anonymization is not applied, the ORF system generates the scores from original speech. Although personal identifiers (e.g., name, roll number, grade) are encrypted, a speaker's identity can be inferred from the voice signal itself. In this setting, the ORF assessment system is reliable. however, without anonymization, privacy is compromised given that the recordings are available to anyone who has access to the system including the teacher, leaving students at risk of identity exposure in the event of a data leakage.

In case 2, when strong anonymization methods are applied, the identity of the speaker is effectively concealed, making it nearly impossible for teachers or the ORF system to infer the student identity from the voice signal. This ensures a high degree of privacy protection. However, such aggressive anonymization often introduces distortions in speech, such as unnatural timbre, distorted prosody, or lower intelligibility. These distortions can compromise the usefulness of the speech data for ORF assessment.

Case 3 represents a balanced anonymization scenario, which preserves the speaker's identity while retaining essential aspects of the speech signal, including linguistic content, prosody, and age-related vocal characteristics of a child's voice. This approach ensures that the speech remains intelligible and natural, allowing accurate ORF assessment while simultaneously protecting the speaker's identity.

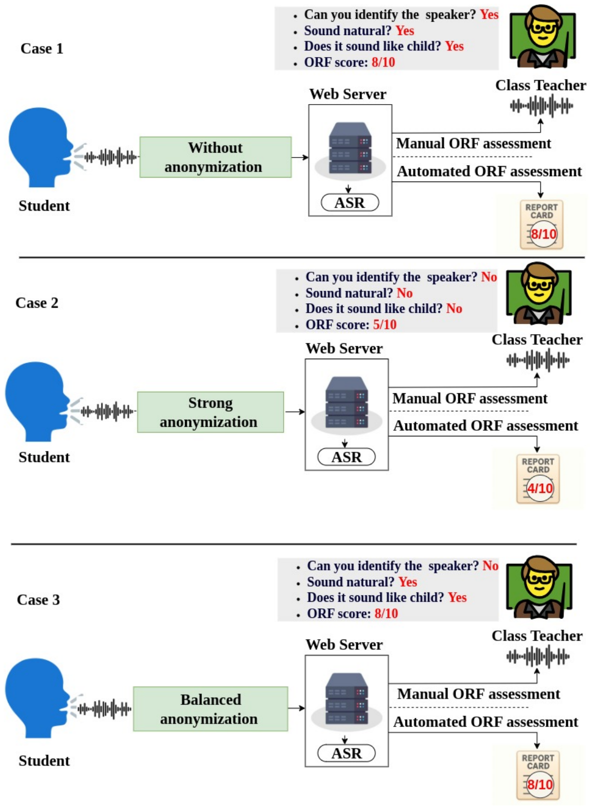


Figure 1: Overview of speaker anonymization in ORF assessment systems.

Achieving a balance between privacy and utility constitutes the core motivation of this work.

Dataset Description

SpeechOcean

The SpeechOcean dataset considered in this work is a subset of the SpeechOcean762 (Zhang et al. 2021) dataset, which contains 5000 English utterances from 250 non-native Mandarin speakers, half of whom are children. Similar to (Kulkarni et al. 2025), we selected a subset of 150 utterances taken from 10 child speakers (15 utterances per speaker) aged 6-10 years. The speakers and utterances match those considered in (Kulkarni et al. 2025). The utterances had durations in the range 2-3 seconds.

MPS

The MPS dataset (Gothi et al. 2024) contains speech utterances recorded in a school environment from children in Grades 3, 4, and 5 corresponding to the age group of 7-11 years. The samples were recorded from 10 different schools across the Indian states of Maharashtra and Goa. Each grade was assigned a unique single story with 2 paragraphs. The children from a specific grade read either one or both the paragraphs assigned to their grade. The speakers represent a cross-section of ORF skills in terms of their word reading errors, pace and prosody. All utterances are transcribed at the word level, carefully preserving any word reading errors, as described by Gothi et al. (2024).

The original dataset has 1600 L2 English utterances (sampled at 16 KHz) from 1110 unique children. However, since the recordings were collected in real world environments, some recordings contained background noise. Hence, we filtered the dataset based on the signal-to-noise ratio estimates and selected 150 utterances for the experiments in this work, each of duration 10-50 seconds, to ensure a balanced comparison with the SpeechOcean dataset. We used the WADA-SNR algorithm (Kim and Stern 2008) to select ”clean” recordings where the minimum estimated local SNR (minimum over a 6 s window), computed at the frame level (6 ms frame) was over 20 dB. We also ensured that the dataset was balanced in terms of the grade and gender of the children. In addition to the clean dataset, we selected a separate set of 60 noisy recordings, with average SNR between 5-10 dB to analyze the impact of noise on the speaker anonymization and the utility. The details of both clean and noisy MPS dataset used for the experimentation are provided in Table 1. The recordings in the SpeechOcean dataset are noise-free and hence not included in the noise robustness analysis.

Speaker Anonymization Methods

In this section, we briefly discuss the working mechanisms of the speaker anonymization methods considered in this paper, namely VTLN (Qian et al. 2018), McAdams coefficient (Patino et al. 2021), and NACLM (Panariello et al. 2023).

Vocal Tract Length Normalization (VTLN)

VTLN is a well known signal processing technique commonly used in voice transformation tasks, such as voice conversion (Sundermann and Ney 2003). It operates by applying a frequency warping function (linear or non-linear) to the spectral envelope of the speech signal, thereby altering the formant frequencies. VTLN simulates the acoustic impact of varying vocal tract lengths, effectively modifying speaker dependent characteristics, which makes VTLN an effective approach for speaker anonymization (Qian et al. 2018).

Several approaches have been studied for the spectral warping (Sundermann and Ney 2003), among which the power transformation is most commonly used. In this paper, we study VTLN with power transformation function based spectral envelope warping and further extend it with pitch shifting. A key element of the proposed method is the analysis-synthesis approach used to extract the envelope and reconstruct the speech after the modification. We use the WORLD vocoder (Morise, Yokomori, and Ozawa 2016) for the analysis-synthesis, which is described below followed by the proposed modifications.

WORLD Vocoder Analysis-Synthesis: VTLN can be effectively implemented using the WORLD vocoder. In the analysis stage, WORLD vocoder decomposes the speech signal into three components: fundamental frequency (F0), spectral envelope (S), and aperiodicity (AP). The F0 contour is extracted with the Harvest algorithm (Morise 2017), the spectral envelope is estimated using the CheapTrick algorithm (Morise 2015), and the aperiodicity is obtained through the D4C algorithm (Morise 2016). In VTLN, the original spectral envelope $S(f)$ is frequency warped according to a warping factor α producing the transformed envelope,

$$S'(f) = S(W(f, \alpha)), \tag{1}$$

where $W(f, \alpha)$ defines the frequency-mapping function. $W(f, \alpha)$ can be either a linear or a non-linear transformation function (Sundermann and Ney 2003). This transformation effectively simulates a change in the speaker’s vocal tract length, thereby modifying perceived speaker characteristics.

During the synthesis stage, the transformed spectral envelope $S'(f)$, along with modified (pitch-shifted) or unmodified F0 contour and AP, is used by WORLD to reconstruct the speech waveform. This step combines all three components to produce a high quality speech signal that reflects the intended vocal tract modifications while maintaining naturalness and intelligibility.

Power Transformation Function: Power transformation based VTLN is a nonlinear frequency warping technique. Given $S(f)$ for a frame of speech (where f denotes the frequency bins), the first step is to normalize the frequency axis as follows,

$$x = \frac{f}{f_{\max}}, \quad x \in [0, 1], \tag{2}$$

where $f_{\max}$ represents the maximum frequency, and x is the normalized frequency. Thereafter, the power transformation is applied to the normalized frequency axis:

$$\begin{aligned} x' &= W(x, \alpha) \\ &= x^{\alpha}, \end{aligned} \tag{3}$$

where α is the warping factor that determines the degree of warping. Since the warping is applied in the normalized frequency domain, the frequency axis is compressed for $\alpha > 1$ causing formants to shift downward, and for $\alpha < 1$ the frequency axis is stretched moving the formants towards higher frequencies. The warped normalized frequencies x' are then mapped back to the original frequency axis to reconstruct the spectral envelope $S'(f)$.

Pitch Shifting: To apply a pitch shift of n semitones, the fundamental frequency F_0 is modified as follows,

$$F_0' = F_0 \cdot 2^{\frac{n}{12}} \quad . \tag{4}$$

The modified frequency F_0' is then used during synthesis to generate pitch modified samples. A positive value of n corresponds to an increase, while a negative value of n corresponds to a decrease in the fundamental frequency of the modified sample.

McAdams Coefficient Method

The McAdams coefficient method (Patino et al. 2021) anonymizes speech by altering its timbre via a nonlinear modification of formant frequencies computed using the source-filter model of speech. In this method, short time linear prediction (LP) analysis is performed on the input signal to decompose it into the LP coefficients and the residual

Table 1: Summary of the SpeechOcean, clean MPS, and noisy MPS datasets

Dataset	Grade	# Utterances	# Unique Speakers	#Spkrs (=1 utt.)	#Spkrs (>1 utts.)
SpeechOcean	–	150	10	–	10
MPS	3	53	38	23	15
	4	43	36	29	7
	5	54	38	22	16
Noisy MPS	3	24	17	10	7
	4	20	16	12	4
	5	16	14	12	2

component. LP coefficients are then transformed into the equivalent all-pole representation and the complex poles ϕ (poles with non-zero imaginary part) are raised to a power α, as given in Eq. (5),

$$\phi' = \phi^{\alpha}, \tag{5}$$

referred to as the McAdams coefficient. This operation shifts the formants, effectively altering the perceived vocal characteristics. In the resynthesis stage, the residual component and the modified LP coefficients are used to obtain the anonymized speech signal.

Neural Audio Codec Language Modeling (NACLM)

NACLM (Panariello et al. 2023) is a deep learning based method that uses a neural audio codec (NAC) combined with a language model for speaker anonymization. Unlike the modular approaches (Meyer et al. 2023; Meyer, Lux, and Vu 2024) which modify the pitch, linguistic features and speaker traits separately, NACLM considers a unified framework to effectively bottleneck speaker-related information using a pool of speaker prompts as additional inputs. It is to be noted that NACLM anonymizes by generating the speech in one of the pseudo speakers voice, which typically corresponds to an adult speaker, and there is no explicit control over the prosodic attributes of the generated output.

Evaluation Methods

In this work, we evaluate the effectiveness of the considered speaker anonymization methods using the objective and subjective measures. The objective evaluation assesses the models in terms of both privacy and utility, while the subjective evaluation considers naturalness, age-group estimation, and speaker identification. Details of the evaluations are described in the following.

Objective Evaluation

To quantify the level of privacy achieved by each method, we use the standard speaker verification task and compute the Equal Error Rate (EER) metric. EER is computed by estimating the cosine similarity between speaker embeddings of anonymized and original speech samples. 192-dimensional speaker embeddings are extracted using a pre-trained ECAPA-TDNN model trained on the VoxCeleb2 dataset (Desplanques, Thienpondt, and Demuynck 2020). Pairs with both samples belonging to one speaker are discriminated from the pairs that comprise two different speakers. For evaluation, each anonymized embedding serves as a trial utterance, and all original (unmodified) utterances are treated as enrollment utterances, simulating an ignorant attacker scenario (Tomashenko et al. 2022). For the MPS and SpeechOcean datasets, the number of same-speaker versus different-speaker pairs are 226 vs. 22,274 and 2,250 vs. 20,250, respectively. For the noisy MPS dataset, the corresponding same- vs. different-speaker pairs are 86 vs. 3,514.

To assess utility, we employ Word Error Rate (WER) (Park, Chen, and Hain 2024) as the primary metric by performing ASR on anonymized speech to evaluate the preservation of linguistic content. We use the Whisper Large-v3 model (Radford et al. 2022), which is trained on multilingual data including English, to obtain the ASR outputs. A lower WER indicates more reliable ORF scores in terms of matching a human rater's estimate of the number of words read correctly by the child. Hence, an anonymization method that does not worsen the WER relative to that achieved on the original utterance is desired.

The proper use of phrase boundaries and prominence are important indicators of reading fluency, with pitch variation being the most important acoustic cue (Sabu and Rao 2021). Therefore, as an additional utility metric, we estimate the normalised F0 contour (cents) across the utterance and measure the degradation due to anonymization by means of the RMSE error, voicing accuracy and the Pearson correlation between the F0 contours of the original and modified utterances. Pitch features are extracted using the Parselmouth library with suitably set analysis parameters for children's speech (Jadoul, Thompson, and de Boer 2018).

Subjective Evaluation

We conducted listening experiments to subjectively evaluate, (i) the naturalness of the anonymized samples, (ii) perceived age-group, and (iii) the speaker anonymization. A total of 12 subjects participated in the evaluation. For the naturalness and age-group tests, we considered three distinct samples per method, forming three different test sets. Each test set included one sample per method, also covering different variations in pitch and warping factors. Participants were assigned one of the three test sets, each containing 20 samples. Subjects were asked to rate the naturalness of each sample on a scale of 1 to 5, with 1 indicating completely unnatural and 5 indicating completely natural. The SpeechOcean and MPS datasets include children aged 6-10 and 7-11 years, and the considered methods shift formant frequencies, leading to a perceived increase or retention of age. Hence, in the age-group test, we asked the subjects to assign the perceived speaker's age in to three categories, below 11 years, between 11-18 years, and older than 18 years (Kulkarni et al. 2025).

In the speaker identification test, a given test item contained an unmodified reference sample and three test samples, one of which is from the same speaker as the reference speaker and two from different speakers, all anonymized using the same method. Three test items were prepared for each anonymization method to better assess the robustness. Subjects were asked to identify the test sample matching the reference speaker, with an additional option of "cannot say" to avoid random guesses under strong anonymization scenarios. The same setup was applied to the original (unmodified) samples.

Results and Discussion

Objective Evaluation

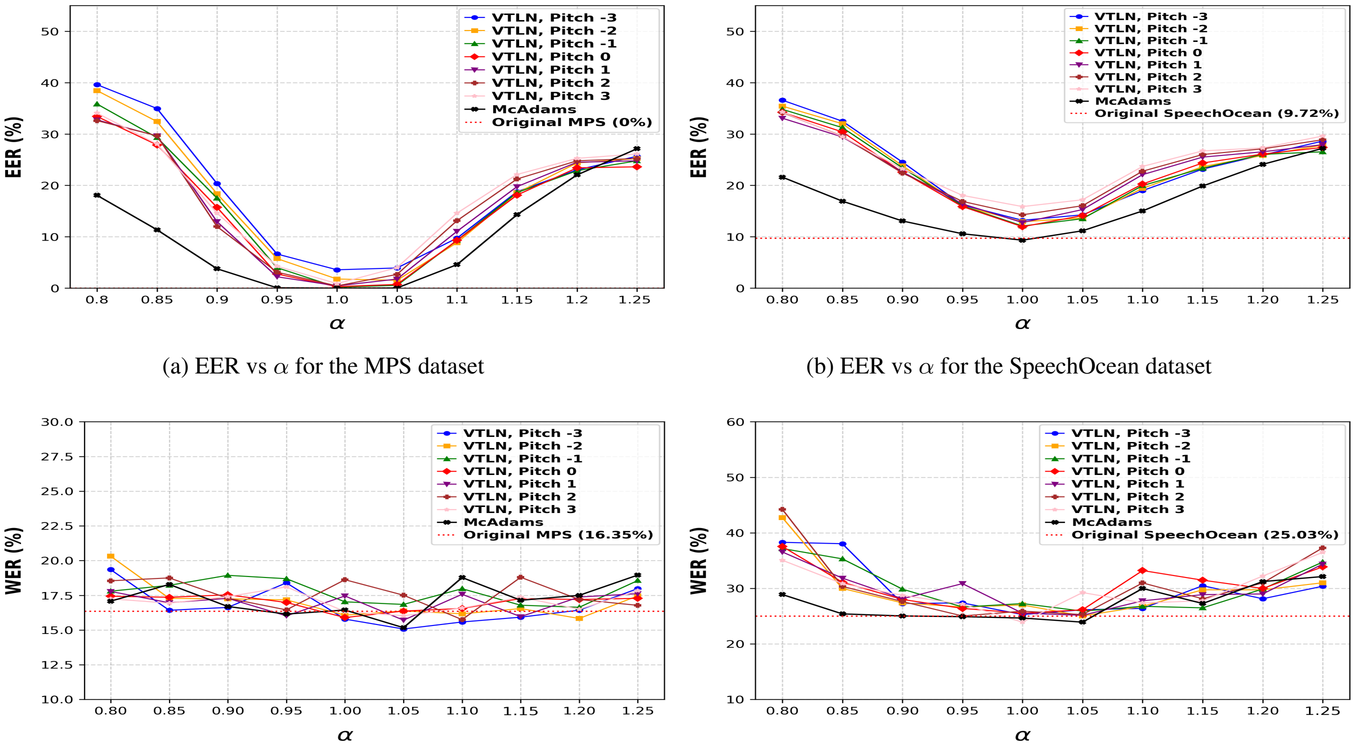
To investigate the impact of pitch shifting and spectral warping on privacy and utility, we conducted experiments on both the MPS and SpeechOcean datasets.

EER/WER Analysis with VTLN: We explored pitch shifts within the semitone range $\{-3, -2, -1, 0, 1, 2, 3\}$ in combination with spectral warping factors varying from 0.80 to 1.25 in steps of 0.05. The parameter ranges were decided based on preliminary experiments, which showed that the methods produce unnatural outputs for warping factors and semitone shifts outside these ranges. For the original (unmodified) speech samples, we observed an EER of 0% for MPS and 9.72% for SpeechOcean. The higher EER in the case of SpeechOcean is attributed to the higher speaker similarity. For both the MPS and SpeechOcean datasets, first, we conducted experiments with no pitch shift (pitch shift = 0) to examine the impact of the spectral warping parameter α on EER. As shown in Figure 2a and Figure 2b, both datasets display a U-shaped trajectory for α values ranging from 0.80 to 1.25, with the highest EER occurring at $\alpha = 0.80$. For $\alpha > 1$ the EER has its maximum at 1.25. However, with informal listerning, we observed that anonymization at $\alpha = 1.25$ produces comparatively more natural speech than at $\alpha = 0.80$. As previously discussed, $\alpha > 1$ decreases the formant center frequencies, producing more adult-like samples, whereas at $\alpha = 0.80$, the samples sound less natural because the formants, which are naturally at higher frequencies for children, are further shifted upward leading to degradation of naturalness. WER estimated across each dataset reveals the similar fluctuating behaviour for both MPS and SpeechOcean with no clear trend except possibly worsening WER at the extremes. This reflects the complex relation between the parameter controlling anonymization and our utility metric of WER.

Next, we analyzed the effect of pitch shift applied in combination with spectral warping, as shown in Figure 2. The overall trend for EER resembles the no pitch-shift case of VTLN across varying α, with significantly higher values observed when $\alpha < 0.90$ or $\alpha > 1.20$. The EER trends reveal two interesting patters. First, when the formant and the pitch are changed in a congruent manner, that is for the combinations with $\alpha > 1$, -ve pitch shifts, and $\alpha < 1$, +ve pitch shifts, the EER remains almost the same as with VTLN using only spectral warping and no pitch shift, indicating that these combinations provide minimal additional anonymization. Second, incongruent changes in the pitch and format frequencies, that is for $\alpha < 1$, -ve pitch shifts, and $\alpha > 1$, +ve pitch shifts, the EER increases, but informal listening showed these combinations produce unnatural outputs. Only the combination ($\alpha > 1$, -ve pitch shifts), which shifts the formants downward and shifts the pitch towards adult pitch range, is found to produce natural speech with minimal distortions and also preserve the spoken content while anonymizing the samples. Apart from EER, WER analysis shown in Figure 2c and Figure 2d indicates that content preservation is better maintained with negative shifts. This may be explained by the increased similarity of the modified voice with adult voices that the ASR has been trained on. For MPS, WER reaches its minimum at –3 semitones in the $\alpha \in [1.0, 1.2]$ range, and even at $\alpha = 1.0$, negative shifts reduce WER compared to the no-shift baseline. Similarly, SpeechOcean shows improved WER for negative shifts, with –3 semitones being optimal. Overall, these findings demonstrate that while $\alpha = 1.20$ provides the best privacy–utility trade-off under no pitch shift, incorporating negative semitone shifts further enhances the balance, with –2 semitones being optimal for MPS and –3 semitones for SpeechOcean.

EER/WER performance of McAdams: Next, we studied the performance of the McAdams Coefficient method as a function of its warping parameter α. Figure 2 shows the EER and WER trends, which closely resemble those of VTLN but with relatively lower values for α. In the case of McAdams method, no further pitch shifting was applied. At $\alpha = 1$, the McAdams EER matches the original datasets, whereas VTLN shows increased EER for SpeechOcean due to moderate post-synthesis distortion. WER results indicate that α values close to 1 lead to low WER (i.e., $\alpha \in [0.95, 1.05]$). Although maximum privacy is achieved at $\alpha = 0.80$ and 1.25, $\alpha = 0.80$ provides the optimal balance between privacy and utility as measured by EER and WER for both MPS and SpeechOcean datasets. Following (Patino et al. 2021), we select the value of α which offers an optimal trade-off between privacy and utility while preserving naturalness, as also confirmed by our experiments at $\alpha = 0.80$.

Privacy-Utility Evaluation across Methods: Based on the optimal parameters obtained from VTLN and McAdams coefficient method, we carried out the objective evaluation to measure prosody preservation. We included NACLM in our comparison to evaluate its performance in terms of privacy and utility relative to the two considered signal-processing-based approaches. NACLM was included in this experiment because it achieved significantly high privacy preservation among the voice-privacy challenge baselines(Tomashenko et al. 2024). The objective evaluation results for all considered methods on the MPS and SpeechOcean datasets are presented in Tables 2. As expected, NACLM provides the strongest privacy protection across both datasets, but at the cost of substantially degraded WER performance and highly distorted prosodic features, as reflected in large pitch RMSE and poor pitch correlation. In contrast, the McAdams method at $\alpha = 0.8$ offers moderate anonymization with only a small increase in WER compared to the original data, while still preserving voicing accuracy and pitch correla-



(a) EER vs α for the MPS dataset

(b) EER vs α for the SpeechOcean dataset

(c) WER vs α for MPS dataset

(d) WER vs α for SpeechOcean dataset

Figure 2: Effect of VTLN warping factor and McAdams coefficient α on EER and WER for the MPS and SpeechOcean datasets.

tion, thus presenting a more balanced privacy-utility trade-off. VTLN demonstrates greater flexibility in tuning the privacy-utility balance: with $\alpha = 1.0$, it preserves utility almost perfectly, whereas at $\alpha = 1.2$, it achieves substantially higher anonymization with only a modest rise in WER. Notably, combining spectral warping at $\alpha = 1.2$ with negative pitch shifts further improves the balance between privacy and utility, with high voicing accuracy and pitch correlation. These findings show that NACLM maximizes privacy but severely degrades intelligibility and prosody, likely because it is not trained on children's speech. In contrast, VTLN with spectral warping and negative pitch shifts, as well as the McAdams method with $\alpha = 0.80$, offer a favorable balance between privacy and speech utility. Additionally, signal processing methods achieve this trade-off in a cost-efficient manner, with lower inference time and without the training overhead required by deep learning approaches[1].

Evaluation for Noisy MPS Speech Samples Based on the optimal McAdams coefficient and VTLN parameters, we tested noisy MPS samples. Since the NACLM method showed the worst utility preserving performance on the clean MPS and SpeechOcean datasets, it was excluded from the noisy-speech evaluation.

Table 3 shows the impact of noise and enhancement on the privacy-utility trade-offs, specifically highlighting the results for the considered methods. Speech enhancement is carried out using the pre-trained DEMUCS model (Défossez, Synnaeve, and Adi 2020), which improves the SNR by 20–30 dB. Raw noisy recordings yield an EER of 0% and WER of 21.43%, higher than the clean MPS WER (16.35%). Under noisy condition, McAdams coefficient method achieves higher anonymization but reduced utility in terms of WER and prosody preservation, whereas VTLN provides slightly lower EER with better WER and prosody preservation. It is observed that Enhancement alone improves WER and preserves prosody. Combining enhancement with McAdams increases anonymization but reduces utility. Whereas, applying VTLN on enhanced data strikes the best balance. Overall, the use of speech enhancement improves both privacy and utility, and we observe that McAdams favors stronger privacy at higher utility cost, while VTLN achieves a better trade-off.

Subjective Evaluation

The subjective evaluation results in Table 4 highlight the influence of anonymization methods on naturalness, perceived age, and speaker identity. As expected, original recordings from both MPS and SpeechOcean datasets achieved the highest naturalness scores, while NACLM received the lowest ratings, indicating clear degradations in speech quality for NACLM. McAdams and VTLN achieved intermediate scores, with VTLN showing slightly higher naturalness for the MPS dataset, while McAdams performed better on SpeechOcean.

For the perceived age-group test, results indicate that the anonymization consistently shifted children's voices toward older groups. NACLM produced the strongest shift, with

[1] Speech samples of the considered methods are available at: https://sandipandhartsk03.github.io/Demo-page-DAPLAB/

Table 2: Evaluation results on MPS and SpeechOcean datasets. 'Org' denotes the original speech. NACLM, McAdams ($\alpha = 0.8$), and VTLN ($\alpha = 1.20$, ps = $-2/-3$) are labeled M1, M2, and M3, respectively, where ps indicates the pitch shift. The best results among the three for each dataset are highlighted in **bold**.

Method	**EER (%)↑**	**WER (%)↓**	**Pitch RMSE (in cent)↓**	**Voicing Accuracy↑**	**Pitch Correlation↑**
MPS					
MPS-org	0.00	16.35	0	100±0.0	1.00±0.0
NACLM (M1)	**47.87**	42.54	426.15±140	75.30±9.93	0.12±0.16
McAdams, $\alpha = 1$	0	16.44	1.50±6.52	99.09±0.03	0.99±0.0
McAdams, $\alpha = 0.8$ (M2)	18.07	17.07	74.22±26	89.11±4.02	0.91±0.25
VTLN, $\alpha = 1$	0.29	15.89	14.84±6.47	98.34±2.02	0.98±0.02
VTLN, $\alpha = 1.20$	23.41	17.19	57.68±25	92.45±1.86	0.95±0.02
VTLN, $\alpha = 1.20, ps = -2$ (M3)	24.36	**15.83**	**55.28±22.15**	**93.44±1.65**	**0.96±0.03**
SpeechOcean					
SpeechOcean-org	9.72	25.03	0	100±0.0	1.00±0.0
NACLM (M1)	**45.83**	50.83	274.82±149.39	77.63±13.08	0.17±0.44
McAdams, $\alpha = 1$	9.33	24.65	0.98±3.03	99.90±0.00	1±0.00
McAdams, $\alpha = 0.8$ (M2)	21.60	28.89	66.38±36.92	90.91±4.37	0.92±0.34
VTLN, $\alpha = 1$	11.95	25.52	12.73±4.76	97.06±4.01	0.98±0.24
VTLN, $\alpha = 1.20$	26.09	29.98	51.82±33.94	93.01±4.29	0.95±0.14
VTLN, $\alpha = 1.20, ps = -3$ (M3)	25.91	**28.12**	**54.68±34**	**92.13±6.92**	**0.94±0.39**

Table 3: EER, WER, and Prosody Metrics for Noisy and Enhanced MPS dataset

Method	**EER (%)↑**	**WER (%)↓**	**Pitch RMSE (in cent)↓**	**Voicing Accuracy↑**	**Pitch Correlation↑**
MPS-org (noisy)	0.00	21.43	0	1	1
McAdams, $\alpha = 0.8$ (M2)	**20.92**	25.42	130±58.47	84.31±4.77	0.65±0.31
VTLN, $\alpha = 1.20$, ps=-2 (M3)	17.39	**22.24**	**101.37±60.80**	**89.10±3.02**	**90.60±0.09**
MPS-org (enhanced)	0.00	19.50	0	1	1
McAdams, $\alpha = 0.8$ (M2)	**23.28**	21.86	88.98±36.87	86.22±3.81	0.75±0.51
VTLN, $\alpha = 1.20$, ps=-2 (M3)	19.77	**19.74**	**51.01±33.44**	**91.11±3.00**	**93.75±0.7**

Table 4: Naturalness, Perceived Age-group, and Speaker Identification results for MPS and SpeechOcean datasets.

Method	**Naturalness↑**	**Perceived Age-group**			**Speaker Identification**		
		<11	**11-18**	**>18**	**Correct**	**Incorrect**	**Can't say**
MPS							
MPS-org	4.27 ± 1.01	8	4	0	36	0	0
NACLM (M1)	2.27 ± 0.79	0	1	11	0	0	36
McAdams, $\alpha = 0.8$ (M2)	3.00 ± 0.77	3	7	2	6	14	16
VTLN, $\alpha = 1.20$, ps=-2 (M3)	**3.18 ± 0.75**	6	5	1	4	14	18
SpeechOcean							
SpeechOcean-org	4.09 ± 1.04	9	3	0	36	0	0
NACLM (M1)	2.91 ± 0.70	0	2	10	0	0	36
McAdams, $\alpha = 0.8$ (M2)	**3.27 ± 0.79**	2	9	1	4	18	14
VTLN, $\alpha = 1.20$, ps=-3 (M3)	3.18 ± 0.75	4	7	1	2	9	25

nearly all samples perceived as > 18 years, while McAdams and VTLN caused moderate shifts toward the 11-18 years range. This indicates that although formant shifting effectively anonymizes speech, it alters the perceived age of children; however, the perceived age remains within the under-18 age group, thereby still fulfilling our intended objective.

Speaker identification results further validate the effect of anonymization on human perception. This subjective evaluation is motivated by the potential unsuitability of the speaker embeddings for children's voices, and the consequent lower reliability of objective EER. Original recordings yielded perfect identification, whereas NACLM achieved full anonymization (100% "can't say"). McAdams and VTLN produced anonymization, with some correct matches but a large share of "can't say" responses. VTLN resulted in fewer correct identifications compared to McAdams, suggesting stronger anonymization though still less effective than the deep learning–based NACLM. In summary, NACLM provides the strongest privacy but degrades both naturalness and age perception, whereas McAdams and VTLN offer a balanced compromise by maintaining naturalness while moderately improving speaker privacy.

Conclusion

In this work, we systematically evaluated the privacy-utility trade-offs of different speaker anonymization strategies for children's speech We found that anonymization using signal processing methods provides a balanced compromise by preserving naturalness and intelligibility in terms of WER and prosody. In contrast, neural approaches like NACLM achieve near-complete identity suppression but at the cost of utility and naturalness. These findings underscore that anonymization is inherently a multi-objective optimization problem, requiring a balance between privacy, utility, and listener acceptability. Despite these insights, some shortcomings were identified including the limited effectiveness of McAdams and VTLN in suppressing speaker identity, with distortions introduced for some parameter settings. Future research should prioritize child specific anonymization methods that retain the acoustic characteristics essential for ORF assessment and preserve the original age group of the child speakers, extend these approaches to multilingual settings to maintain language-dependent prosodic and phonetic cues, and advance next-generation techniques that are secure and child-sensitive.

Acknowledgment

We thank Kulkarni et al. for sharing the SpeechOcean speaker list used in (Kulkarni et al. 2025), and we also thank all the participants of our subjective evaluation experiments.

References

Arasteh, S. T.; Arias-Vergara, T.; Pérez-Toro, P. A.; Weise, T.; Packhäuser, K.; Schuster, M.; Noeth, E.; Maier, A.; and Yang, S. H. 2024. Addressing challenges in speaker anonymization to maintain utility while ensuring privacy of pathological speech. *Communications Medicine*, 4(1): 182.

Bailly, G.; Godde, E.; Piat-Marchand, A.-L.; and Bosse, M.-L. 2022. Automatic assessment of oral readings of young pupils. *Speech Communication*, 138: 67–79.

Desplanques, B.; Thienpondt, J.; and Demuynck, K. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech*.

Défossez, A.; Synnaeve, G.; and Adi, Y. 2020. Real Time Speech Enhancement in the Waveform Domain. In *Proceedings of Interspeech*, 3291–3295.

Evanini, K.; and Wang, X. 2013. Automated speech scoring for non-native middle school students with multiple task types. In *Proceedings of Interspeech*, 2435–2439.

Gothi, R.; Kumar, R.; Pereira, M.; Nayak, N.; and Rao, P. 2024. A Dataset and Two-pass System for Reading Miscue Detection. In *Proceedings of Interspeech*, 4014–4018.

Jadoul, Y.; Thompson, B.; and de Boer, B. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71: 1–15.

Kim, C.; and Stern, R. M. 2008. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. In *Interspeech 2008*, 2598–2601.

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.

Kulkarni, A.; Teixeira, F.; Hermann, E.; Rolland, T.; Trancoso, I.; and Doss, M. M. 2025. Children's Voice Privacy: First Steps And Emerging Challenges. *arXiv preprint arXiv:2506.00100*.

Meyer, S.; Lux, F.; Koch, J.; Denisov, P.; Tilli, P.; and Vu, N. T. 2023. Prosody Is Not Identity: A Speaker Anonymization Approach Using Prosody Cloning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Meyer, S.; Lux, F.; and Vu, N. T. 2024. Probing the Feasibility of Multilingual Speaker Anonymization. In *Proceedings of Interspeech*, 4448–4452.

Morise, M. 2015. CheapTrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Communication*, 67: 1–7.

Morise, M. 2016. D4C, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, 84: 57–65.

Morise, M. 2017. Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals. In *Interspeech 2017*, 2321–2325.

Morise, M.; Yokomori, F.; and Ozawa, K. 2016. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, E99.D(7): 1877–1884.

Panariello, M.; Nespoli, F.; Todisco, M.; and Evans, N. W. D. 2023. Speaker Anonymization Using Neural Audio Codec Language Models. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4725–4729.

Park, C.; Chen, M.; and Hain, T. 2024. Automatic Speech Recognition System-Independent Word Error Rate Estimation. *ArXiv*, abs/2404.16743.

Patino, J.; Tomashenko, N. A.; Todisco, M.; Nautsch, A.; and Evans, N. W. D. 2021. Speaker Anonymisation Using the McAdams Coefficient. In *Proceedings of Interspeech*, 1099–1103.

Qian, J.; Du, H.; Hou, J.; Chen, L.; Jung, T.; and Li, X.-Y. 2018. Hidebehind: Enjoy Voice Input with Voiceprint Unclonability and Anonymity. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*, SenSys '18, 82–94. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359528.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning*.

Sabu, K.; and Rao, P. 2021. Prosodic event detection in children's read speech. *Computer Speech & Language*, 68: 101200.

Sundermann, D.; and Ney, H. 2003. VTLN-based voice conversion. In *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No.03EX795)*, 556–559.

Tomashenko, N.; Miao, X.; Champion, P.; Meyer, S.; Wang, X.; Vincent, E.; Panariello, M.; Evans, N.; Yamagishi, J.; and Todisco, M. 2024. The VoicePrivacy 2024 Challenge Evaluation Plan.

Tomashenko, N. A.; Wang, X.; Miao, X.; Nourtel, H.; Champion, P.; Todisco, M.; Vincent, E.; Evans, N. W. D.; Yamagishi, J.; and Bonastre, J.-F. 2022. The VoicePrivacy 2022 Challenge Evaluation Plan. *ArXiv*, abs/2203.12468.

Yoo, I.-C.; Lee, K.; Leem, S.; Oh, H.; Ko, B.; and Yook, D. 2020. Speaker Anonymization for Personal Information Protection Using Voice Conversion Techniques. *IEEE Access*, 8: 198637–198645.

Zhang, J.; Zhang, Z.; Wang, Y.; Yan, Z.; Song, Q.; Huang, Y.; Li, K.; Povey, D.; and Wang, Y. 2021. speechocean762: An Open-Source Non-native English Speech Corpus For Pronunciation Assessment. In *Proceedings of Interspeech*.