scientific reports

Check for updates

OPEN

fNIRS experimental study on the impact of AI-synthesized familiar voices on brain neural responses

Weijia Zhang[1,2,3,4], Jiaju Li[1], Luyang Ji[1], Xue Cheng[1], Dongqin Sun[1], Yanrong Jiang[1], Feiyu Chen[1], Yiduo Zhou[1], Calvin Choi[5], Hao Cheng[6] & Shaomin Cai[7]✉

With the advancement of artificial intelligence (AI) speech synthesis technology, its application in personalized voice services and its potential role in emotional comfort have become research focal points. This study aims to explore the impact of AI-synthesized familiar and unfamiliar voices on neural responses in the brain. We utilized the GPT-SoVITS project to synthesize three types of voices: a female voice, a sweet female voice, and a maternal voice, all reading the same text. Using functional near-infrared spectroscopy (fNIRS), we monitored the changes in blood oxygen levels in the prefrontal cortex and temporal cortex of participants during listening, assessing brain activation. The experimental results showed that the AI-synthesized maternal voice significantly activated the participants' prefrontal and temporal cortices. Combined with participants' feedback, the activation of these areas may reflect multidimensional features of voice familiarity processing, including emotion, memory, and cognitive function. This finding reveals the potential applications of AI voice technology in enhancing mental health and user experience.

Keywords Artificial intelligence, Human-computer interaction, Voice synthesis, Social impact of synthetic speech, fNIRS

With the rapid development of artificial intelligence (AI) technology, significant progress has been made in speech synthesis in recent years, with the gap between AI-synthesized voices and human voices continuously narrowing[1,2]. At the same time, numerous studies have started to focus on voice personalization by learning individual voice samples to mimic the voice of a specific person[3,4]. In particular, the emergence of Few-Shot Text-to-Speech (FSTS) technology marks a breakthrough in the field of personalized speech synthesis[5]. FSTS technology learns and captures the unique timbral characteristics of a specific speaker from a small number of voice samples, enabling the generation of personalized voices even in data-scarce conditions.

AI-powered personalized speech synthesis could impact individuals in various ways. With the advancement of speech synthesis technology, fraudulent activities involving the imitation of personal voices have emerged. Scammers can record or synthesize the voice of a target to impersonate a trusted individual or authority figure, thereby gaining access to sensitive information or funds. One study discussed how AI-generated tremulous voices affect users' perceptions, emotional experiences, and subsequent behaviors, mentioning that fraudsters have used voice cloning technology to impersonate the CEO of a British energy company[6]. Existing research has begun to explore the extent to which audio deepfakes can deceive vulnerable users[7].

However, we should also focus on the positive aspects of this technology. In the design of voice assistants for mental health, humanization, emotional expression, and conversational usage are all crucial[8]. AI-driven realistic speech synthesis holds significant potential in the field of speech-based interactions for mental health and emotional support. As a companion robot for emotional support, using voice interaction to alleviate loneliness and social isolation among the elderly has been widely studied[9,10]. However, the connection between AI-synthesized voices and human emotional responses has not been fully explored. In particular, there is a lack of direct neuroscientific evidence regarding the impact of AI-synthesized familiar voices (e.g., the voices of loved ones) on emotional and brain responses.

[1]Department of Mathematica and Information Sciences, Shaoxing University, Shaoxing, China. [2]Key Laboratory of Artificial Intelligence Multi-dimension Applications, Shaoxing University, Shaoxing, China. [3]Integrated Circuit Innovation Class, Shaoxing University, Shaoxing, China. [4]Visiting Scholar, Department of AOP Physics, University of Oxford, Oxford, UK. [5]Oxford Industrial Holding Group, Kowloon, Hong Kong. [6]Zhejiang Chengshi Metaverse Technology Co., Ltd, Hangzhou, China. [7]School of Medicine, Shaoxing University, Shaoxing, China. ✉email: usxmedlab@yeah.net

Extensive research has been conducted on the recognition of familiar versus unfamiliar voices[11]. Therefore, we hypothesize that there should also be differences between AI-synthesized familiar and unfamiliar voices. Based on research on brain regions involved in speech processing and the distinction between familiar and unfamiliar sounds, we propose the following hypothesis: the neural response in relevant brain regions will be stronger when listening to the familiar voice of a loved one compared to when listening to an unfamiliar voice. Furthermore, the response activated by familiar voices will influence the individual's emotional state, thereby eliciting certain emotional reactions.

Studies have shown that speech processing involves several key brain regions, primarily located in the temporal and prefrontal areas. Mathiak et al.[12] found through fMRI studies that the temporal lobe plays a crucial role in distinguishing speaker identity, especially in regions located in the left posterior superior temporal gyrus (STG) and the right temporal pole. Nakamura et al.[13], using PET scans, also demonstrated that when distinguishing familiar (e.g., friends or one's own voice) from unfamiliar voices, there was significant activation in the left frontal pole and right temporal pole. Further research[14], using MEG technology, confirmed this phenomenon, showing notable activity in the right superior temporal sulcus (STS) 200 milliseconds after performing voice recognition tasks, while speech tasks mainly activated the left STG region. Meanwhile, the inferior frontal gyrus (IFG) in the prefrontal cortex is involved in both speech production and processing[15]. The anterior temporal and frontal lobes also play important roles in identity recognition, participating in the processing of voice identity, and may be involved in the multimodal integration of voice and facial information[16].

The prefrontal cortex (PFC) plays a critical role in cognitive control and emotional regulation[17]. Monitoring activity in this region can reveal the impact of voice interactions on emotional regulation strategies and emotional expression. The temporal cortex (TC) is involved in the processing of language and emotional memory[18]. Monitoring its activity can explore how voice interactions influence the retrieval of emotional memories and the expression of emotional content. Therefore, we can monitor brain activation in the prefrontal cortex and temporal cortex while listening to AI-synthesized voices to roughly assess whether they can evoke emotional responses in humans. The locations of the prefrontal cortex and temporal cortex are shown in Fig. 1.

Functional Near-infrared Spectroscopy (fNIRS) is a non-invasive brain imaging technique that monitors brain activity in real-time by measuring changes in cortical blood oxygen levels[19]. fNIRS works by emitting near-infrared light and detecting the reflected signals that pass through the scalp and skull. It can differentiate the relative concentration changes of oxygenated hemoglobin (Oxy-Hb) and deoxygenated hemoglobin (Deoxy-Hb), inferring the neural activation state of a specific brain region. When fNIRS detects an increase in Oxy-Hb concentration, the concentration of Deoxy-Hb typically decreases, indicating that neural activity in that region has increased, meaning the brain area is "activated"[20]. This technology has the advantage of high temporal resolution. Additionally, fNIRS offers key benefits in studying auditory mechanisms because, compared to fMRI (Functional Magnetic Resonance Imaging), it operates more quietly[21], making it particularly suitable for research on cognitive processes such as speech processing and emotional responses. Due to its portability and minimal interference, fNIRS has become an essential tool in neuroscience research.

Compared to traditional speech synthesis studies, this research utilizes fNIRS technology to directly measure changes in blood oxygen levels in the prefrontal and temporal regions when processing AI-synthesized voices of strangers and familiar voices of loved ones, reflecting the degree of brain activation. This represents an innovative approach. Additionally, using fNIRS, the study aims to explore the neural responses of the brain to familiar and unfamiliar AI-synthesized voices, specifically investigating whether familiar voices (such as those of loved ones) can significantly activate brain regions in the prefrontal and temporal areas, thereby influencing emotional responses. If the neural response in the prefrontal cortex and temporal cortex is stronger when hearing familiar voices of loved ones compared to unfamiliar voices, and participants exhibit emotional reactions, it would suggest that AI-synthesized familiar voices can influence the brain's neural responses and evoke voice familiarity processing. This would provide a solid foundation for the application of AI voice technology in psychological health interventions and enhancing user experiences.

Materials and methods

This study will utilize the GPT-SoVITS (Generative Pre-trained Transformer-SoftVC VITS Singing Voice Conversion) model for voice synthesis, with the specific steps illustrated in Fig. 2.

First, audio samples will be collected for training and processed accordingly. After audio processing is completed, the model training phase will begin, utilizing the GPT-SoVITS model for speech generation. The synthesized speech will be used for subsequent experiments.

Second, synthesized speech will be presented as stimuli to participants while their brain activity data is recorded using fNIRS.

Finally, the collected fNIRS data will undergo preprocessing, followed by block averaging and channel averaging, leading to statistical analysis to evaluate the impact of speech on brain activity.

Speech synthesis

Audio sample collection

Before the experiment, we first collected audio samples from the mothers of all participants. To cover the vocal characteristics of the participants' mothers as comprehensively as possible and reduce interference from differences in audio content, we used GPT to generate a text that would take approximately one minute to read. The participants' mothers were asked to read this text, and the resulting audio sample was used for synthesizing the mother's voice using GPT-SoVITS. The other two stranger voice samples were sourced from publicly available free materials on the internet, consisting of a middle-aged woman's voice and a sweet female voice.

Both the middle-aged woman's voice and the sweet female voice are in standard Mandarin. When requesting audio samples from the participants' mothers, we specifically asked them to use Mandarin. However, since the

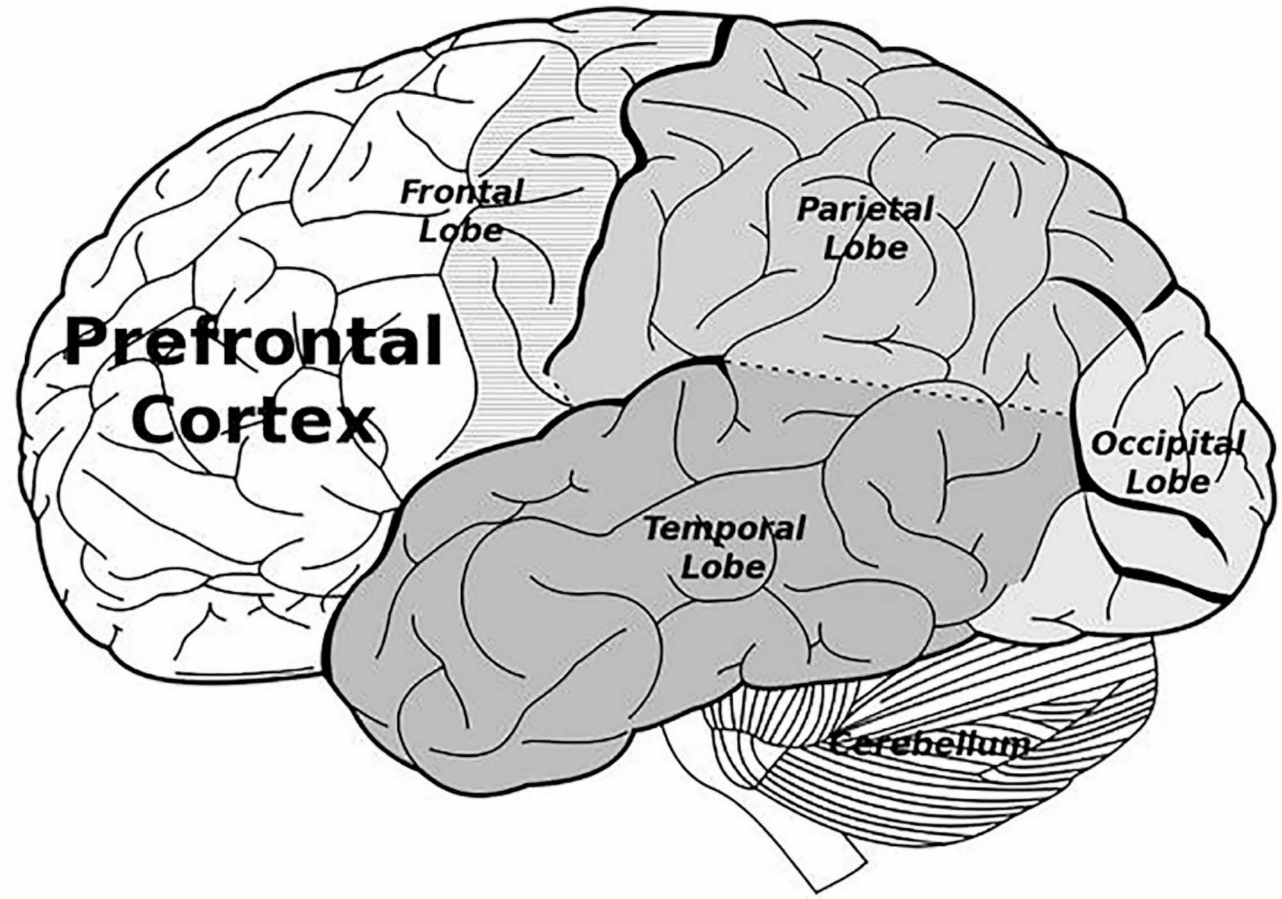

Fig. 1. Schematic diagram of the locations of the Prefrontal Cortex and Temporal Cortex. The Prefrontal Cortex is located at the front of the brain and is involved in cognitive control and emotional regulation, while the Temporal Cortex is situated below the Prefrontal Cortex and is responsible for processing language and emotional memory. The division of these regions is based on the classification of brain lobes. The cerebral cortex is a thin layer of neural tissue that covers the entire surface of the brain, approximately 2–4 mm thick, and contains the brain's gray matter. The Prefrontal Cortex is part of the frontal lobe, as indicated in the diagram. The Temporal region corresponds to the Temporal Cortex.

participants' mothers belong to a certain age group, their Mandarin inevitably carries a regional dialect accent. Nevertheless, this accent is widely understandable to people from various regions and can be categorized as a common dialect-accented Mandarin.

The middle-aged woman's voice refers to the voice of a female who sounds around 40 years old, with a relatively stable pitch, typically in the mid-to-low frequency range, and a rich, full tone. The sweet female voice, on the other hand, resembles the voice of a woman around 20 years old, with a higher pitch, usually in the mid-to-high frequency range, and a clear, bright, and light tone. We provided specific audio examples for reference.

In terms of age, the middle-aged woman's voice is more similar to the participants' mothers' voices, while it forms a clear age contrast when compared to the sweet female voice.

Audio processing and model training

This study uses the GPT-SoVITS project for voice synthesis, a deep learning-based few-shot voice synthesis model that improves upon the original SoVITS (SoftVC VITS Singing Voice Conversion) model[22] and integrates the GPT model. The GPT-SoVITS model has been publicly released and can be accessed on the open-source code repository GitHub.

SoVITS adopts an improved encoder, the SoftVC encoder, which can more accurately capture the speaker's voice characteristics, including timbre, pitch, and prosody, enhancing the model's conversion performance across different speakers. GPT-SoVITS inherits and uses this encoder. It also introduces a residual quantization layer, which receives input features from the encoder and transforms them into more compact and model-friendly semantic encoding features. These features contain the vocal characteristics of the reference audio, allowing the model to better mimic the reference voice when generating speech. During the inference stage, the autoregressive module is used to gradually generate new speech segments based on the semantic encoding features of the reference audio, ensuring that the generated speech is consistent with the reference in terms of

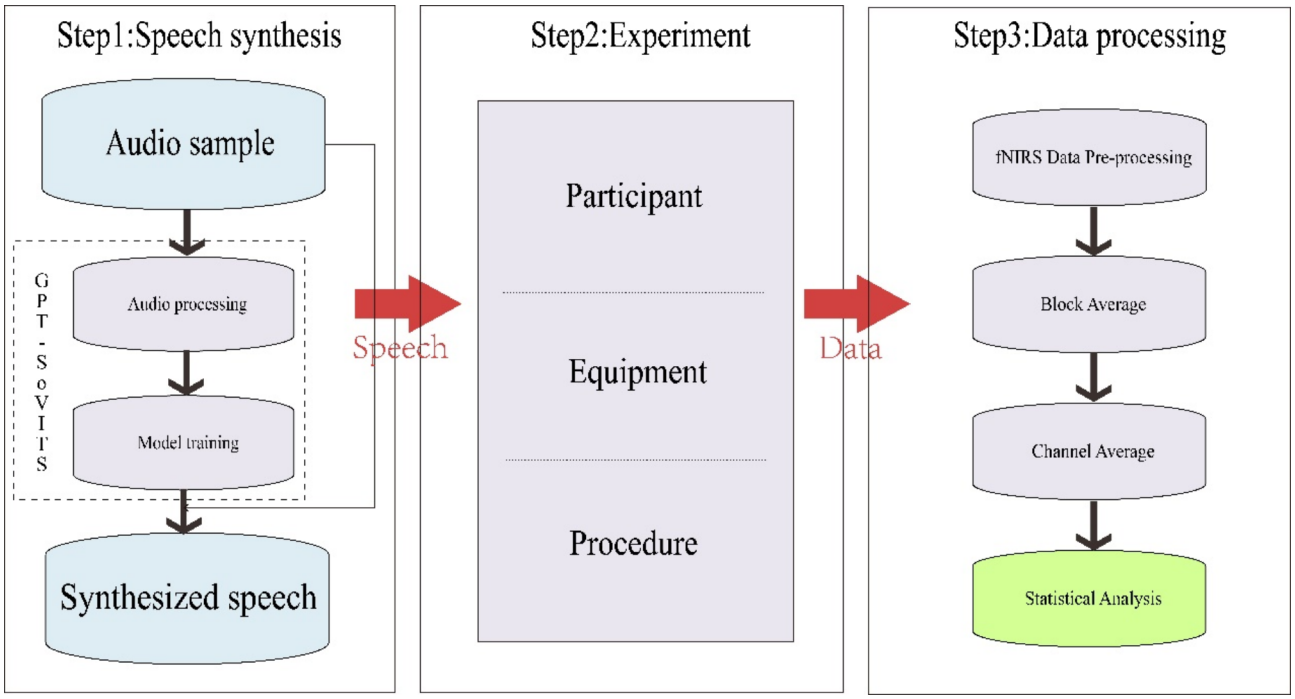

Fig. 2. Overall framework diagram.

vocal features. This allows the speech synthesized by GPT-SoVITS to significantly surpass typical AI-generated speech projects in terms of naturalness, prosody, rhythm, pronunciation imitation, and emotional expression. Additionally, it can efficiently generate speech with a specific speaker's timbre using a very small amount of reference audio. Specific technical details can be found in Appendix 1, and the process is shown in Fig. 3.

Since the middle-aged woman's voice and the sweet female voice come from a high-quality database, no additional noise reduction processing was required. However, the recording quality of the participant's mother varied, so noise reduction was applied to improve the audio quality and ensure comparability in sound quality among the three samples. Afterward, all three voices were segmented, annotated with ASR (Automatic Speech Recognition) text, and used for model training for speech synthesis.

Synthesis and segmentation

The trained SoVITS model and GPT-like model are selected, and a segment of 3 to 10 s of audio is extracted from the audio samples as the reference audio. The voice synthesis will refer to the tone and speed of this audio segment, combined with the pre-set text content to generate the corresponding speech. The pre-set text for this experiment is selected from Long Yingtai's " *The Letters of André*," primarily consisting of a letter written by a mother to her son. We use three types of audio samples to generate the same content: AI-generated middle-aged woman's voice, sweet girl's voice, and mother's voice. The specific process is illustrated in Fig. 4, with a total of 25 synthesized materials from different participants' mothers.

Audio synthesized by the model was segmented into 25-second clips using a Python program for use in subsequent experiments.

Experiment

Participants

A total of 25 student volunteers aged between 20 and 25 years were recruited for this study, including 12 males and 13 females. All volunteers met the following criteria: (a) normal hearing, including self-assessment and a simple conversation test; (b) no neurological disorders affecting the experimental results; (c) no brain tumors and/or brain structural abnormalities caused by trauma; (d) a Montreal Cognitive Assessment (MoCA) score of ≥ 28.

All participants signed informed consent forms and provided their consent prior to each test. This study was approved by the Shaoxing People's Hospital Ethics Review Committee. The experimental procedures adhered to the requirements of the medical ethics committee and complied with the ethical standards set forth in the 1975 Declaration of Helsinki (revised in 2008).

Among the participants, 20 took part in Experiment 1 (a comparison of AI-generated female voices and AI-generated maternal voices), while 13 participated in Experiment 2 (a comparison of AI-generated sweet female voices and AI-generated maternal voices). For details, see Sect. "Experimental procedure".

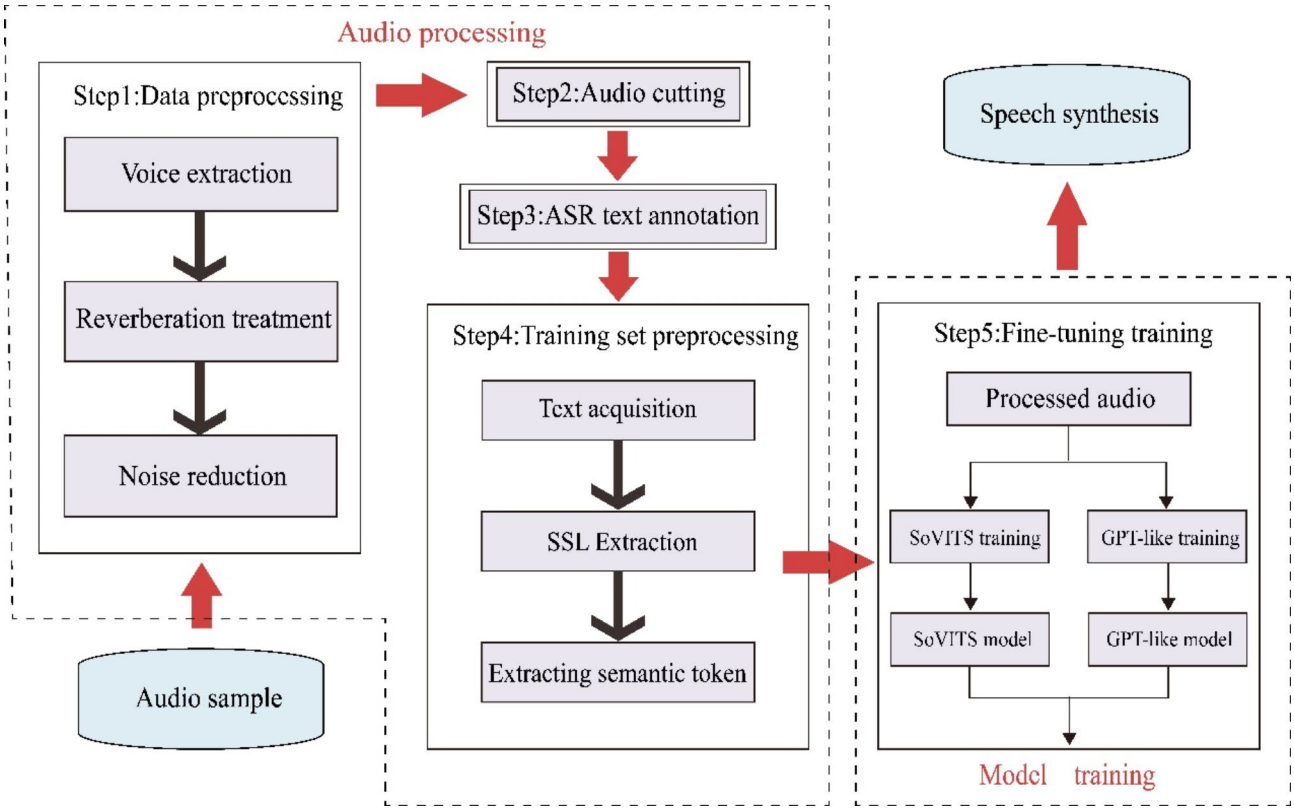

Fig. 3. Flowchart of audio processing and model training.

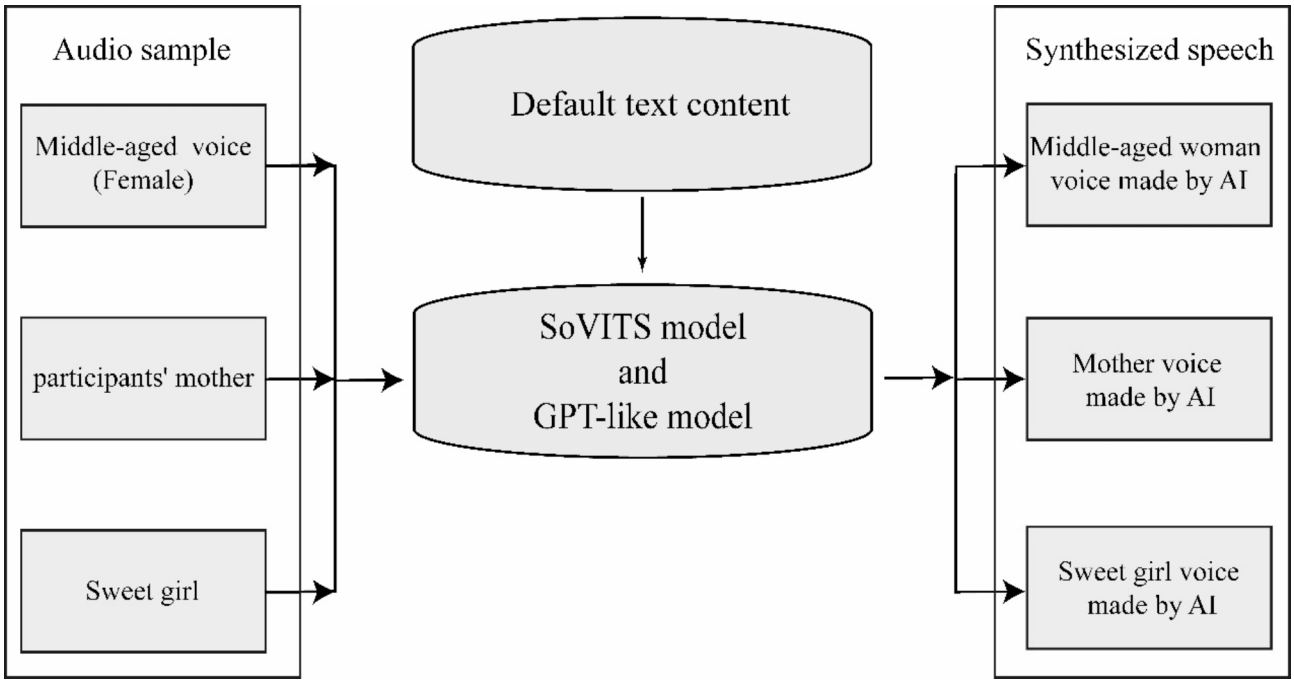

Fig. 4. Experimental voice synthesis workflow diagram.

Equipment

The fNIRS sampling device used in this experiment was the NirSmart (Danyang Huichuang Medical Equipment Co., Ltd., China). This device records near-infrared spectroscopy signals using wavelengths of 760 nm and 850 nm, with a frequency of 11 Hz, and has a probe spacing of 3 cm. According to the modified Beer-Lambert law, the device can continuously measure and record the concentration changes of oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) in participants' brains during task performance.

Following the internationally accepted 10/20 electrode placement system, the experiment utilized 10 near-infrared light source probes and 8 detector probes, forming a total of 22 effective channels. These channels were primarily distributed across the temporal and prefrontal regions. To ensure the accuracy of the experimental data, a flexible head-mounted fixture was used to maintain a consistent distance between the emitters and the scalp. Figure 5 shows the two-dimensional and three-dimensional arrangements of the functional near-infrared spectroscopy probe array.

Experimental procedure

The voices used in this experiment are all AI-generated. For simplicity, the description "AI-generated" may be omitted in the following text.

Experiment 1: The voice of an unfamiliar middle-aged woman serves as the control group, while the voice of the experimenter's mother serves as the experimental group.

Experiment 2: The voice of the unfamiliar sweet young woman served as the control group, while the voice of the experimenter's mother served as the experimental group.

The audio text content for both the experimental and control groups is identical; specific details can be found in Sect. "Synthesis and segmentation".

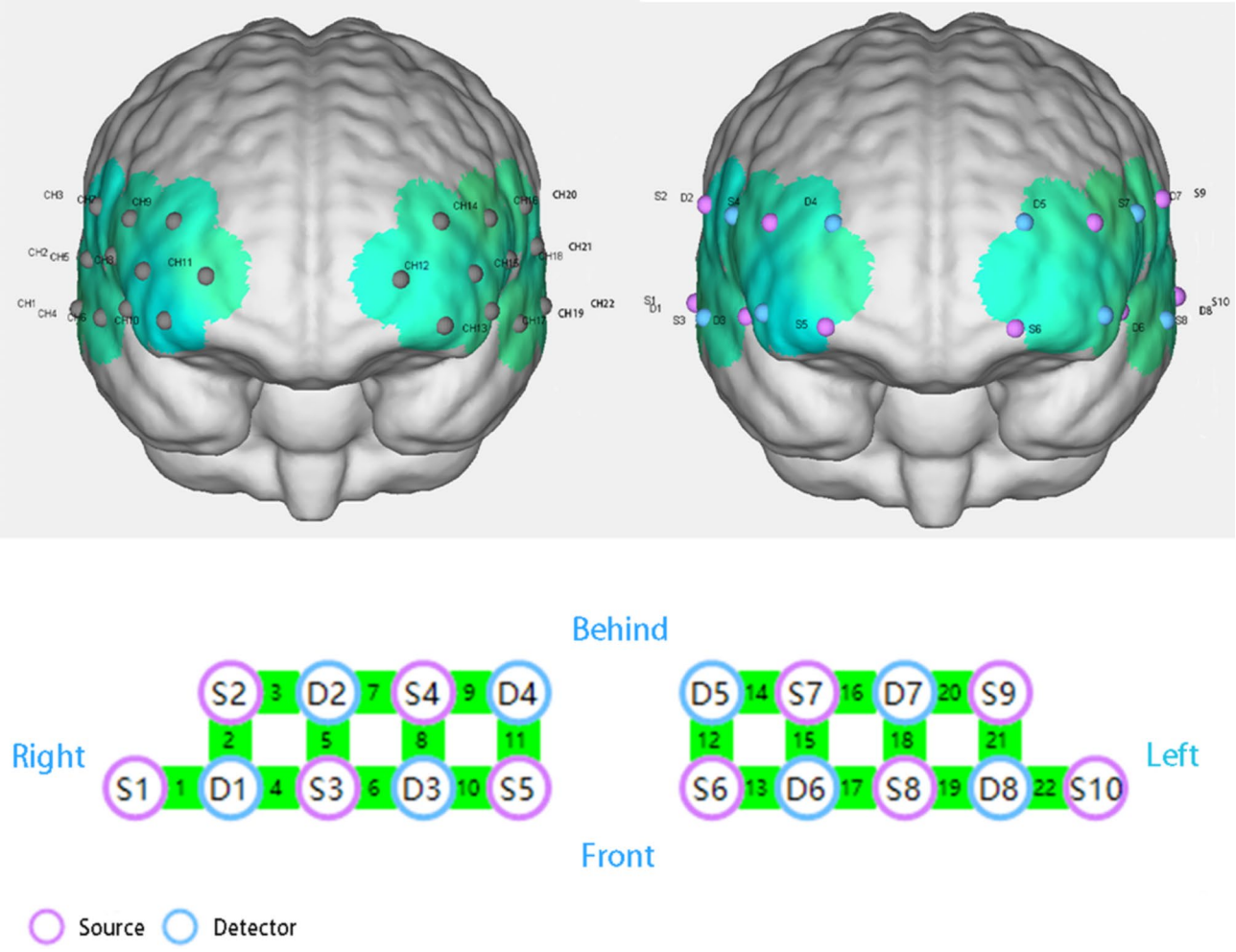

Fig. 5. fNIRS Probe Array Diagram. The left brain diagram marks the positions corresponding to 22 channels, while the right brain diagram indicates the locations of the 10 emitter probes and 8 receiver probes. The lower diagram is a 2D illustration of the channels. Channels CH3, CH5-CH18, and CH20 mainly cover the prefrontal cortex, while CH1, CH2, CH4, CH19, CH21, and CH22 primarily cover the temporal cortex. Data on the brain regions covered by these channels must be exported from the NirSpark software. See Sect. "Block average and channel average" for details.

This experiment uses AI-synthesized speech to provide audio stimuli to the participants, while fNIRS equipment is used to measure and record the changes in blood oxygen concentration during the experiment. The specific experimental setup is shown in Fig. 6.

Experimental procedure

The experiment consists of two main parts: rest state and task state. Each task cycle includes 15 s of rest followed by 25 s of task. Before the experiment begins, participants enter a resting state, serving as the baseline for experimental data. The fNIRS device records data starting from zero, with the baseline value being the blood oxygen concentration measured during the 2 s prior to the start of the experiment. The task cycle is arranged as follows: each cycle consists of 25 s of the unfamiliar voice task, followed by 15 s of rest, then 25 s of the familiar voice task, and another 15 s of rest. The entire experiment includes 10 such cycle repetitions. To ensure the accuracy of the experimental data, the rest periods are designed both to meet the block averaging requirements of the experiment and to provide participants with sufficient rest to prevent fatigue and potential sequence effects. The experimental task is shown in Fig. 7.

Since the experiment involves auditory stimuli, it must be conducted in a quiet environment to prevent interference from external noise. Before the test, participants are instructed to close their eyes and remain still to avoid movement-related disturbances. During the experiment, they are asked to focus on listening to the different audio stimuli presented, ensuring that they maintain their full attention. During rest periods, participants should relax in order to allow their blood oxygen levels to recover quickly. It is essential to ensure that participants fully understand the experimental requirements and are prepared for the experiment.

At the beginning of the experiment, the experimenter first reminds the participants of the relevant details of the experimental process and instructs them to close their eyes and enter a resting state in preparation for listening to the audio. Once the experimenter confirms that the participant has entered a resting state, the formal experiment begins, following the task flow outlined in Fig. 7. Throughout the experiment, participants are only required to listen to the audio, with no additional tasks. The experimenter ensures that the environment remains

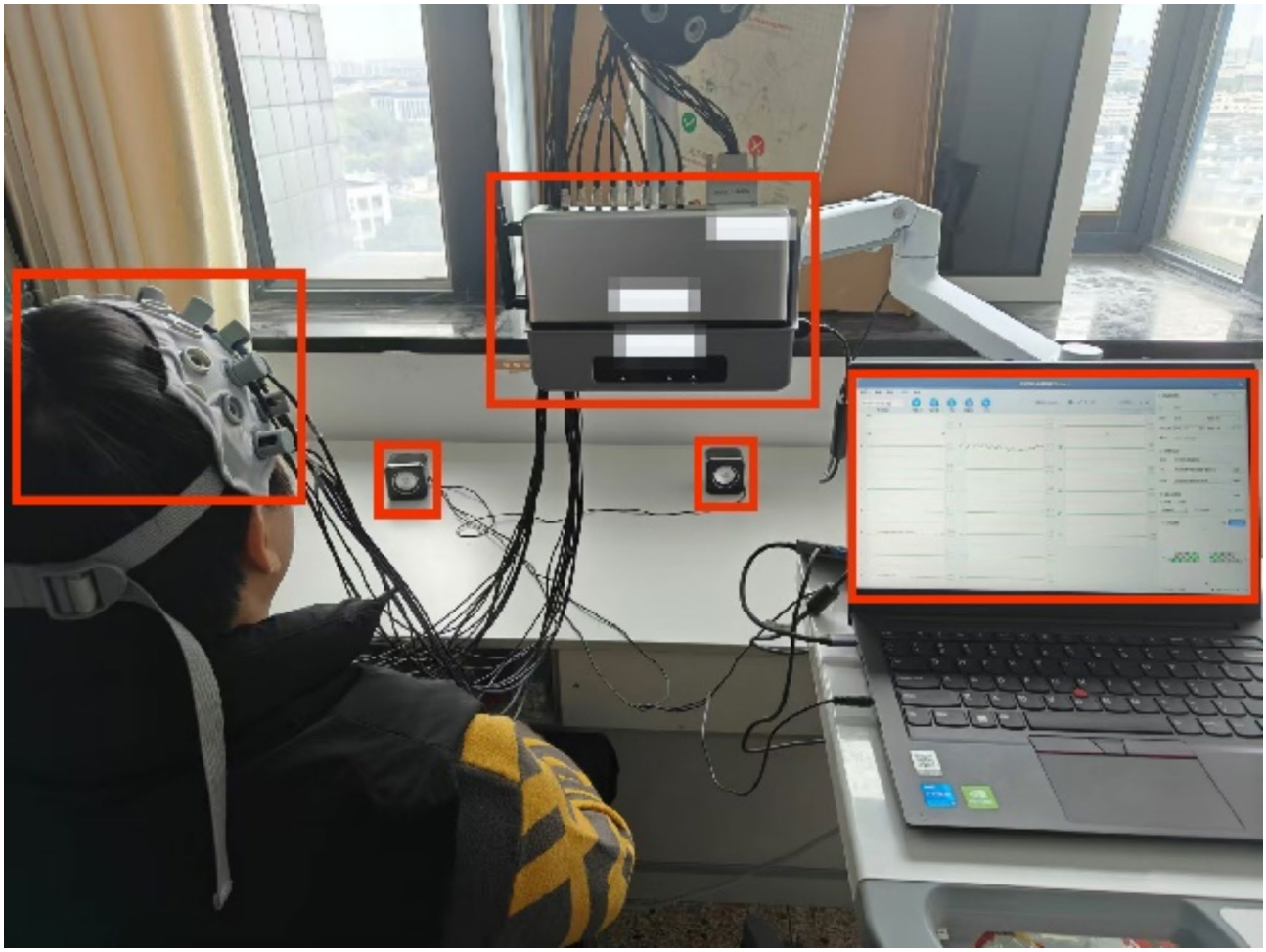

Fig. 6. Experimental Setup. Participants wore head covers equipped with fNIRS probes (red box, left) to measure brain activity. The signals were transmitted via optical fibers to the fNIRS processing unit (center), where they were processed and sent to a connected computer for visualization (red box, right). Audio stimuli were delivered through a speaker (red box, bottom center) positioned in front of the participant.

Fig. 7. Experimental Task Flowchart. Task 1 involves AI-generated unfamiliar voices, while Task 2 involves AI-generated familiar voices. The unfamiliar voices include the female voice from Experiment 1 and the sweet female voice from Experiment 2.

quiet and free from distractions. After the experiment concludes, participants are informed about the survey they need to complete. The survey focuses on recalling any emotional reactions, scene memories, or other unique experiences they had while listening to the AI-generated voice of their mother. The content of the survey can be found in Appendix 2.

Data processing

fNIRS data pre-processing

The data pre-processing was performed using NirSpark software (Danyang Huichuang Medical Equipment Co., Ltd., China). The near-infrared spectroscopy signals contain instrument noise, experimental noise, and physiological noise. Instrument noise mainly arises from environmental interference or signals from the device itself, while experimental noise is primarily caused by motion artifacts from the participant's head movement during the testing process. Physiological noise consists of external physiological signals collected while monitoring brain activity, which can interfere with the fNIRS signals.

To minimize the impact of these noises, the raw signal data collected by the fNIRS device was processed accordingly using the software. During the pre-processing, motion artifacts appear as spikes or abrupt changes caused by relative sliding between the scalp and the probes. To correct these artifacts, the motion standard deviation and spline interpolation methods were used. Studies have proven that this method is both accurate and effective[23], and it demonstrates the effectiveness of preprocessing. The NirSpark software is equipped with algorithms that identify artifacts based on input parameters and automatically retrieve potential artifact intervals in the data. A sliding time window of 0.5 s was used to check the time intervals of the detected artifacts, and spline interpolation was applied to correct the data. The default standard values provided by the software were used. According to previous studies, in order to remove irrelevant low-frequency drifts, filter out high-frequency noise introduced by the equipment, and eliminate physiological noise caused by heartbeat or respiration, a bandpass filter with a range of 0.01 Hz to 0.2 Hz was applied to the data[24,25]. The Hemo module in the software is designed to convert the obtained signals into hemoglobin concentration change images, transforming the raw light intensity data from fNIRS into changes in oxygenated hemoglobin concentration. All differential pathlength factor (DPF) values are set to the default value of 6.0, which is derived from previous studies[26].

At the same time, in the subsequent analysis, we used Oxy-Hb as our primary indicator, as the Oxy-Hb signal typically exhibits a better signal-to-noise ratio than Deoxy-Hb[27].

Block average and channel average

We analyzed the average concentration of oxygenated hemoglobin during the 25-second task period of listening to the voices and compared it with the baseline period to assess the changes in Oxy-Hb concentration between the task and the baseline. Since the software automatically sets the average value during the baseline period to 0, we only needed to analyze the difference between the Oxy-Hb concentration values and the 0 value.

The NirSpark software utilizes the international 10/20 system and predefined anatomical templates to accurately map the channel positions of the fNIRS signals based on Brodmann areas, assisting researchers in determining the brain regions covered by each channel. In this study, the 22 channels covered different functional areas. The channel configuration data was exported from the NirSpark software, as detailed in Table 1.

The proportions in the table represent the relative ratio of each channel covering different areas among all the regions it passes through. Generally, when the ratio exceeds 0.5, a region is considered the primary coverage area. For ease of analysis, we categorized these coverage areas based on broader brain lobe divisions to support subsequent data operations and result interpretation.

Statistical analysis

The schematic diagram of the changes in oxygenated hemoglobin concentration was created using GraphPad Prism software (version 9.5; GraphPad Software, San Diego, USA). First, signals from each brain region under each task condition were extracted, and the average change in oxygenated hemoglobin concentration (ΔHbO) was calculated. The results are presented as mean ± standard error of the mean (SEM), with statistical significance set at $p < 0.05$.

In this study, to investigate the changes in blood oxygen concentration in two brain regions under different task conditions, a Linear Mixed Effects Model (LMM) was used to analyze the fNIRS data. The analysis aimed

Channel	Region (Brodmann areas)	Region (lobar division)	Proportion
CH1	Middle Temporal Gyrus	Temporal Lobe	0.9662
CH2	Superior Temporal Gyrus	Temporal Lobe	0.7596
CH3	Pre-Motor and Supplementary Motor Cortex	Prefrontal Cortex	0.6046
CH4	Middle Temporal Gyrus	Temporal Lobe	0.7793
CH5	Pars Triangularis (Broca's Area)	Prefrontal Cortex	0.4702 (Combined with Opercular area for total of 0.7084)
CH6	Pars Triangularis (Broca's Area)	Prefrontal Cortex	0.4141 (Combined with Inferior Frontal Gyrus for total of 0.6666)
CH7	Pars Triangularis (Broca's Area)	Prefrontal Cortex	1
CH8	Dorsolateral Prefrontal Cortex	Prefrontal Cortex	0.6784
CH9	Dorsolateral Prefrontal Cortex	Prefrontal Cortex	0.8525
CH10	Inferior Frontal Gyrus	Prefrontal Cortex	0.4435 (Combined with Dorsolateral Prefrontal Cortex for total of 0.8826)
CH11	Frontopolar Area	Prefrontal Cortex	0.637
CH12	Frontopolar Area	Prefrontal Cortex	0.5514
CH13	Inferior Frontal Gyrus	Prefrontal Cortex	0.4619 (Combined with Dorsolateral Prefrontal Cortex for total of 0.7309)
CH14	Dorsolateral Prefrontal Cortex	Prefrontal Cortex	0.8554
CH15	Dorsolateral Prefrontal Cortex	Prefrontal Cortex	0.7707
CH16	Pars Triangularis (Broca's Area)	Prefrontal Cortex	1
CH17	Dorsolateral Prefrontal Cortex	Prefrontal Cortex	0.3574 (Combined with Pars Triangularis for total of 0.6996)
CH18	Pars Triangularis (Broca's Area)	Prefrontal Cortex	0.6084
CH19	Middle Temporal Gyrus	Temporal Lobe	0.7828
CH20	Pre-Motor and Supplementary Motor Cortex	Prefrontal Cortex	0.6064
CH21	Superior Temporal Gyrus	Temporal Lobe	0.4512 (Combined with Pre-Motor and Supplementary Motor Cortex for total of 0.6192)
CH22	Middle Temporal Gyrus	Temporal Lobe	0.9228

Table 1. Brain regions covered by channels and their proportions.

to evaluate the impact of task sound type and brain region on brain activity. The Linear Mixed Effects Model analysis was conducted using IBM SPSS Statistics (version 30.0.0.0; IBM Corp., Armonk, USA). The LMM is a statistical method suited for handling repeated-measures data, capable of considering both fixed and random effects. This allows for effective control of individual differences, thereby improving the accuracy and robustness of the analysis results.

Specifically, the fixed effects in the model include the task sound type (such as AI-generated mother's voice, middle-aged woman's voice, sweet young woman's voice), brain regions (such as temporal lobe and prefrontal lobe), and their interaction effects. The random effects part accounts for individual differences among participants, incorporating them into the analysis to eliminate biases between different participants. By using this modeling approach, we are able to test the main effects of task sound type and brain region, as well as their interactions, to gain a deeper understanding of how these factors influence brain activity.

To ensure the model's fit and reliability, we calculated various information criteria, including the −2 log likelihood, Akaike Information Criterion (AIC), corrected Akaike Information Criterion (AICC), consistent Akaike Information Criterion (CAIC), and Schwarz Bayesian Criterion (BIC). Additionally, significance tests for both fixed and random effects further validated the model's effectiveness and robustness. To verify the fit of the mixed effects model, covariance parameter estimation analysis was also conducted. Through these analyses, this study was able to precisely assess the impact of task sound type and brain region on brain oxygenation changes across different task conditions.

Results

fNIRS data

In Experiment 1, participants listened to AI-generated maternal voices and AI-generated female voices. The fNIRS data, after processing, were analyzed for the average values and their differences, as shown in Fig. 8.

In Experiment 1, the fNIRS data showed that the average ΔHbO for the maternal voice condition was 0.008703, which was higher than the average ΔHbO for the female voice condition, -0.01329. The difference in means between the two voice types was 0.02199, with the differences observed in specific brain regions, including the frontal lobe and temporal lobe. The mean differences for these regions were 0.01968 for the frontal lobe and 0.02431 for the temporal lobe.These findings suggest that there is a significant difference in brain activation between the maternal voice and the female voice. Specifically, the AI-generated maternal voice elicited higher brain activation levels than the AI-generated female voice.

In Experiment 2, participants listened to AI-generated maternal voices and AI-generated sweet female voices. As shown in Fig. 9, the processed fNIRS data reveal the average ΔHbO values and their differences for these two voice types.

In Experiment 2, the fNIRS data showed that the average ΔHbO for the maternal voice condition was 0.01641, significantly higher than the ΔHbO for the sweet female voice condition, which was −0.006584. The difference in means between the two voice types was 0.02300, with the differences observed in specific brain regions. The mean differences in the frontal lobe and temporal lobe were 0.01946 and 0.02653, respectively.

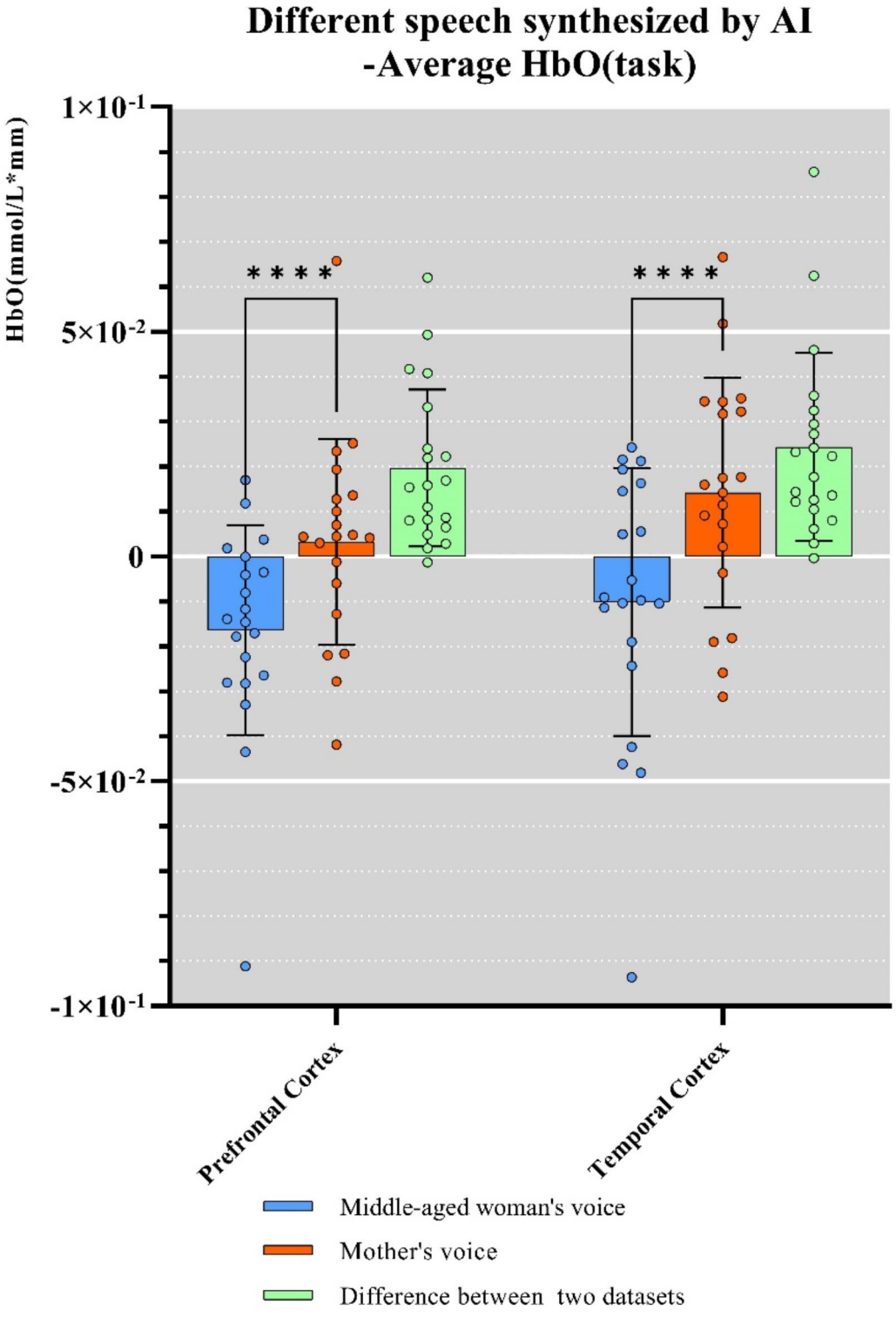

Fig. 8. Comparison of the fNIRS data from participants under two different AI-generated voices in Experiment 1 after averaging, along with their differences (*$p<0.05$; **$p<0.01$; ***$p<0.001$; ****$p<0.0001$).

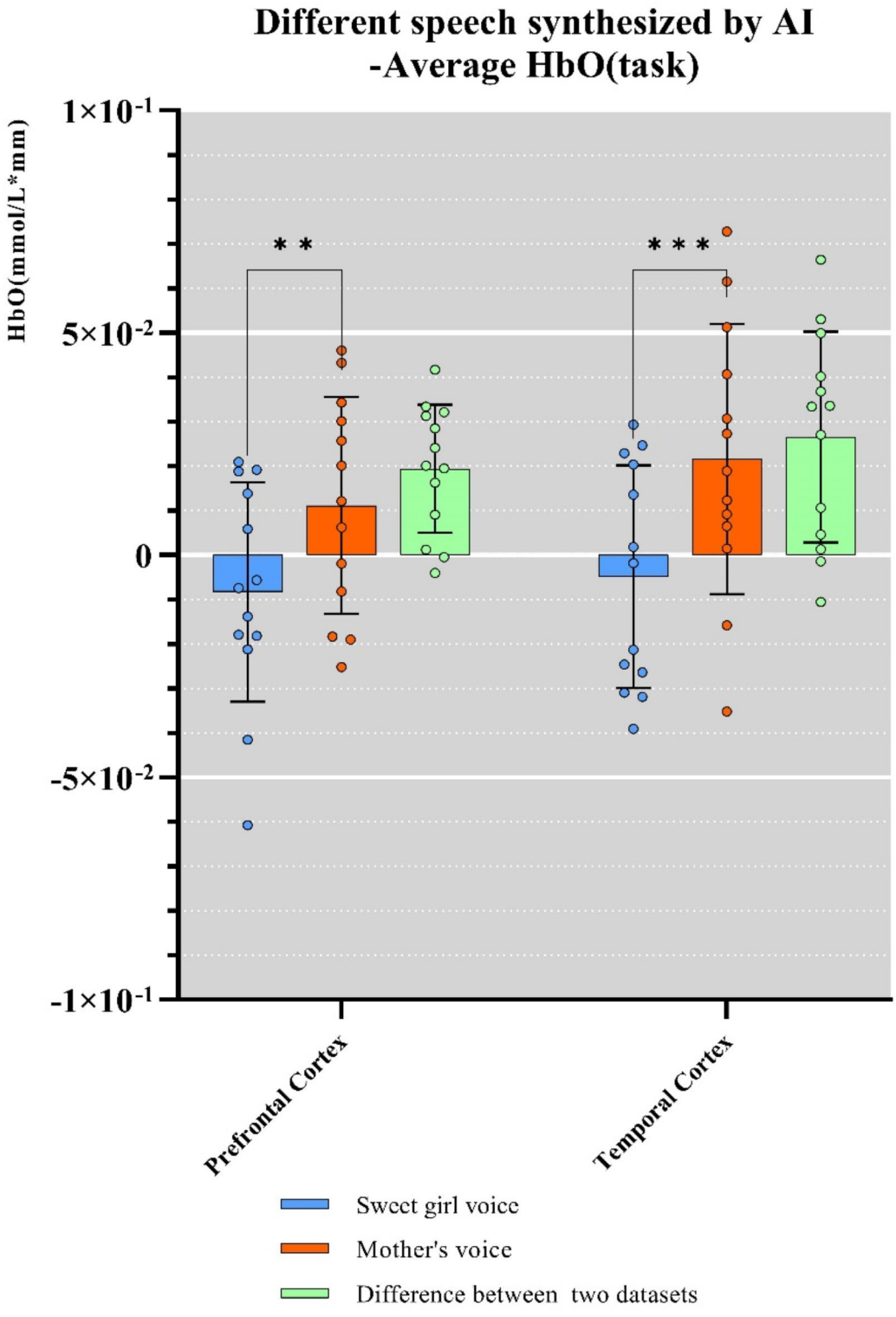

Fig. 9. Comparison of the fNIRS data from participants under two different AI-generated voices in Experiment 2 after averaging, along with their differences (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$).

Group	Factor	Numerator degrees of freedom	Denominator degrees of freedom	F-Value	*p*-Value	Significance
Experiment 1	Intercept	1	19.239	0.222	0.643	Not Significant
Experiment 1	Task Voice Type	1	52.714	41.062	< 0.001	Significant
Experiment 1	Brain Region	1	52.714	6.364	0.015	Significant
Experiment 1	Task Voice Type × Brain Region	1	52.714	0.456	0.503	Not Significant
Experiment 2	Intercept	1	12.963	0.576	0.461	Not Significant
Experiment 2	Task Voice Type	1	23.803	29.459	< 0.001	Significant
Experiment 2	Brain Region	1	23.803	2.682	0.115	Not Significant
Experiment 2	Task Voice Type × Brain Region	1	23.803	0.696	0.412	Not Significant

Table 2. Linear mixed effects model analysis results - fixed effects test.

Group	−2 Restricted Log-Likelihood	AIC	AICC	BIC	CAIC
Experiment 1	−370.992	−360.992	−360.135	−344.338	−349.338
Experiment 2	−237.367	−227.367	−225.938	−213.011	−218.011

Table 3. Mixed effects model fit and information criteria.

These results also indicate that there is a significant difference in brain activation between the mother's voice and the sweet young woman's voice. Specifically, the AI-generated mother's voice induced higher brain activation levels than the AI-generated sweet young woman's voice, which is consistent with the results from Experiment 1.

LMM analysis results

To further assess the impact of task voice type and brain regions on brain activity, a mixed-effects model was used in this study. Table 2 presents the fixed effects test results for the task voice type, brain region, and their interaction effects on the fNIRS data in Experiments 1 and 2. By comparing the F-values and p-values of different effects, we can examine the significant impact of each factor on changes in brain activity. The specific results are shown in Table 2:

In Experiment 1, the main effect of task voice type ($F = 41.062$, $p < 0.001$) and brain region effect ($F = 6.364$, $p = 0.015$) were both significant, indicating that different task voice types and brain regions had a significant impact on brain activity. However, the interaction between task voice type and brain region was not significant ($F = 0.456$, $p = 0.503$), meaning that the combination of different task voice types and brain regions did not significantly alter the brain's response to activity. In Experiment 2, the main effect of task voice type was also significant ($F = 29.459$, $p < 0.001$), while the main effect of brain region was not significant ($F = 2.682$, $p = 0.115$), and the interaction between task voice type and brain region did not reach significance ($F = 0.696$, $p = 0.412$).

To assess the fit of the mixed-effects model, this study calculated several information criteria, including the −2 log-likelihood, Akaike Information Criterion (AIC), corrected Akaike Information Criterion (AICC), consistent Akaike Information Criterion (CAIC), and Schwarz Bayesian Criterion (BIC). We have listed the goodness-of-fit information for Experiment 1 and Experiment 2 to help evaluate the model's fit and the impact of different information criteria on model selection, as shown in Table 3.

Table 3 presents the goodness-of-fit metrics for the models in Experiment 1 and Experiment 2. The values of the information criteria for both experiments are relatively low, indicating that the mixed-effects model fits the data well. In both experiments, the changes in AIC and BIC values suggest that the model in Experiment 1 fits the data slightly better than in Experiment 2, but overall, the model shows good fit and high reliability. These information criteria further support the model's effectiveness in data fitting and provide a solid foundation for subsequent analysis.

To further validate the random effects of the model, we present the estimated values of the covariance parameters in Experiment 1 and Experiment 2. The covariance parameters reflect the contribution of individual differences between subjects to the variance of the fNIRS data. The estimation of the random effects helps evaluate the stability and reliability of the model. The specific results are shown in Table 4.

Table 4 shows the estimated covariance parameters for the random effects part of the mixed-effects model. The estimated covariance values for the experiments indicate that individual differences between subjects contributed minimally to the data variation. Specifically, in both experiments, the variance estimate for the subject intercept was zero, and the covariance estimates for repeated measurements were also very small. This suggests that individual differences between subjects had a minimal impact on data variation in both experiments. The high stability of the model further validates the good fit and high reproducibility of the mixed-effects model.

Discussions

Firstly, in this experiment, participants' psychological states (such as mental stress, sleep quality, etc.) were not specifically measured or controlled. However, since the experiment was conducted continuously in a quiet environment, the participants' states were relatively stable throughout the experiment, and it can be assumed that these factors were relatively controlled during the experiment. Participants were asked to rest and clear their

Group	Parameter	Estimate	Standard Error
Experiment 1	Task Voice Type = AI Synthesized Women's Voice × Brain Region = Temporal Lobe	0	0
Experiment 1	Task Voice Type = AI Synthesized Women's Voice × Brain Region = Frontal Lobe	0	6.49E−05
Experiment 1	Task Voice Type = AI Synthesized Mother's Voice × Brain Region = Temporal Lobe	0	0
Experiment 1	Task Voice Type = AI Synthesized Mother's Voice × Brain Region = Frontal Lobe	0	9.60E−05
Experiment 1	Intercept (Subjects)	0	0
Experiment 2	Repeated Measures: Task Voice Type = Mother's Voice × Brain Region = Temporal Lobe	0.001	0
Experiment 2	Repeated Measures: Task Voice Type = Mother's Voice × Brain Region = Frontal Lobe	9.43E-05	6.42E−05
Experiment 2	Repeated Measures: Task Voice Type = Sweet Female Voice × Brain Region = Temporal Lobe	0	7.06E−05
Experiment 2	Repeated Measures: Task Voice Type = Sweet Female Voice × Brain Region = Frontal Lobe	0	7.60E−05
Experiment 2	Intercept (Subjects)	0	0

Table 4. Covariance parameter estimates.

minds before the experiment and to remain still and focus their attention during the experiment to minimize potential effects of mental stress or fatigue on the results.

Secondly, since the tasks in the experimental procedure were conducted in a fixed order, with unfamiliar voice tasks followed by familiar voice tasks, the order effect might have influenced the participants' psychological states. However, we minimized this effect by arranging the experimental procedure reasonably and providing appropriate rest periods. During the pilot phase of the experiment, we found that familiar maternal voices induced higher brain activation levels, so we placed them later in the fixed sequence. Additionally, each task state allowed for a 15-second rest period. Despite these precautions, there might still be some order effects. In the sweet female voice group (Experiment 2), the mean and difference values of blood oxygen concentration were higher compared to the female voice group (Experiment 1). The sweet female voice typically has a higher pitch, with audio in the mid-to-high frequency range, and it is clear, bright, and light in quality. Compared to the voice of middle-aged women, it may more easily capture the participants' attention, thus triggering stronger brain activity, reflected in the increased blood oxygen levels. This phenomenon could be a manifestation of an order effect, but it does not affect the overall results of the experiment. This is because we are comparing the mean differences between unfamiliar and familiar voices, and the overall increase in blood oxygen concentration does not affect the core results required for the experiment.

Furthermore, this study predicts that familiar voices of loved ones in AI-generated speech will activate neural responses in the brain, leading to increased activity in corresponding brain regions and influencing emotional responses. In Experiment 1, the mean differences between the two voices in the frontal lobe and temporal lobe were 0.01968 and 0.02431, respectively. In Experiment 2, the mean differences between the two voices in the frontal lobe and temporal lobe were 0.01946 and 0.02653, respectively. In both Experiment 1 and Experiment 2, the mean differences were higher than the mean values during the speech tasks. The results indicate that in the frontal lobe and temporal lobe regions, the changes in oxygenated hemoglobin concentration caused by AI-generated maternal voices were significantly higher than those caused by other AI-generated unfamiliar voices, specifically the voices of AI-generated unfamiliar women and sweet female voices. Therefore, even AI-generated voices of familiar loved ones can effectively activate neural responses in the brain.

Furthermore, through the analysis of the linear mixed-effects model, the impact of task voice types on the brain in both Experiment 1 and Experiment 2 was found to be significant, with the advantage of maternal voices in brain activation being confirmed. However, in Experiment 1, the main effect of different brain regions was significant, while in Experiment 2, the main effect of different brain regions was not significant, with p-values of 0.015 and 0.115, respectively. In Experiment 1, the F-value for brain region was 6.364, while in Experiment 2, the F-value for brain region was 2.682, suggesting that the brain region effect in Experiment 2 was not as pronounced as in Experiment 1. This could be related to the sweet female voice used in Experiment 2 and its voice characteristics, leading to differences in how activation effects were expressed in different brain regions. Participants might have had a more consistent response to the sweet female voice, while their response to the maternal voice might have varied, reducing the statistical effect of brain region responses. As previously mentioned in the discussion on sequence effects, the mean and difference values for oxygenated hemoglobin content were slightly higher for the sweet female voice compared to the women's voice group, indicating that the sweet female voice was indeed more stimulating than the women's voice. Neither the brain region effect nor the interaction effect between task voice type and brain region showed statistically significant interactions, indicating that the response differences across brain regions to the task voices were small, and the effects of different voice types on brain activation were independent. The linear mixed-effects model provided a quantitative evaluation of brain responses, confirming the key role of task voice types in brain activation and providing theoretical support for the application of AI-generated voices in emotional research.

Additionally, during the pre-experiment, participants generally reported that when listening to AI-synthesized maternal voices, they would recall images related to their mothers. Therefore, we designed a questionnaire to record these memories. In contrast, for unfamiliar voices, participants typically focused more on the content of the text being read, without generating significant associations or memories, so we did not design a questionnaire to record these responses. Specific feedback from the formal experiment is listed in Appendix 2. The memories recalled were mostly related to the participants' mothers and were typically unconscious associations and recollections during the task, though the specific scenes and content varied from person to person. Since the prefrontal cortex and temporal cortex play important roles in emotional responses, we hypothesize that familiar voices have a stronger impact on people's emotional responses. Of course, the activation of the prefrontal and temporal cortices may also have other cognitive explanations. Activity in the prefrontal cortex might be associated with attention regulation, memory retrieval, and cognitive conflict processing. When hearing a familiar voice, individuals may allocate more cognitive resources to processing that voice, especially related to emotional memories or novelty processing, which could lead to an enhanced response in the prefrontal cortex. The temporal cortex, on the other hand, is closely related to speech processing, semantic understanding, and the retrieval of emotional memories. Familiar voices may activate emotional memories stored in the brain, thereby enhancing activity in the temporal regions. Therefore, the activation of the prefrontal and temporal cortices may reflect the multidimensional nature of voice familiarity processing, including emotion, memory, and cognitive function, rather than being a simple manifestation of emotional responses.

Finally, to eliminate the acoustic variability of AI-generated voices, we conducted an acoustic analysis. Through linear regression analysis, we compared the spectral differences between the original recordings of different participants and the AI-generated voices, as well as the differences in brain activity under various sound task conditions, to determine whether the acoustic variability of AI-generated voices affected the changes in blood oxygen concentration during the experiment. We used Mel-Cepstral Distortion (MCD) to quantify the similarity between synthesized and natural speech. The specific introduction of MCD and the correlation analysis with the linear regression model are provided in Appendix 3. The results from the supplementary Tables S1, S2, S3, and S4 indicate that there is no significant linear relationship between fNIRS data differences and MCD values, as both the regression analysis and ANOVA results fail to show significant correlation or variance changes. The conclusion drawn was that the acoustic variability of AI-generated voices did not affect the changes in blood oxygen concentration during the experiment.

In conclusion, the analysis suggests that AI-generated familiar voices of loved ones are effective in activating neural responses in the brain and eliciting emotional reactions, although the intensity and consistency of individual responses may vary. At the same time, the findings of this study provide a theoretical basis for the application of AI technology in the field of emotional support, which could have a positive impact on product evaluation.

There are some limitations in the current study. First, the sample size included in this study is relatively small, which may affect its accuracy. Age, as one of the factors influencing the generalizability of user data, was primarily between 20 and 27 years in the participants recruited for this study, which may have some impact on the results. Additionally, in the preprocessing of near-infrared spectroscopy data, subjectivity in identifying artifacts may influence the results.

Secondly, due to the limitations of the fNIRS equipment used in this study, we were unable to provide highly precise localization of specific brain structures. Ideally, MRI (Magnetic Resonance Imaging) should be used to precisely localize brain regions for each participant. However, due to experimental constraints, we were unable to perform MRI co-registration for each participant to accurately confirm the precise structural locations in the brain. Therefore, regional analyses based on Brodmann areas (such as IFG, STG, etc.) may have some limitations. Additionally, our equipment is not a full-cover system and does not cover all brain regions associated with the processing of familiar sounds, making regional analysis potentially incomplete. To address this issue, we chose a lobe-based analysis approach, aiming to analyze broader brain regions, thus reducing the impact of individual brain structural differences on the results and ensuring the accuracy of the conclusions.

Finally, we chose to use a fast speech synthesis model primarily to achieve more efficient data collection. While it maintained relatively high clarity and naturalness in terms of audio quality, participants could still distinguish between synthetic and natural voices both in spectral analysis and subjective perception. However, this difference did not undermine the core finding of the study, which is the significant impact of AI-synthesized familiar voices on brain responses. This finding is consistent with previous research on familiar sounds in the context of natural voice tasks. However, a recent functional magnetic resonance study[28] suggests that there may be differences in how natural and artificial sounds are processed. Therefore, we cannot determine whether the observed brain responses are similar to the reaction mechanisms triggered by natural sounds, but we can say that they reflect neural responses related to the processing of familiar sounds. This may suggest that, while the response differences between artificial and natural sounds may gradually narrow with the ongoing progress of speech synthesis technology, human voices and emotional expression are highly complex and subtle, and AI speech technology may not fully replicate or replace these human traits. This shift will have profound implications for the application of voice-based artificial intelligence technology, but it does not affect the broad application potential of AI speech technology in many domains.

Conclusion

This study delved into the impact of AI-synthesized voices on brain activation using fNIRS equipment, with a particular focus on the significance of AI-synthesized familiar voices in neural responses related to speech processing in the prefrontal cortex and temporal cortex. The results indicate that AI-synthesized familiar voices, specifically the voice of a mother, significantly enhanced neural activity in the prefrontal and temporal regions. This suggests that AI-synthesized familiar voices can influence the brain's neural responses and trigger voice

familiarity processing. This finding not only confirms the crucial role of voice familiarity in sound processing but also analyzes how AI-synthesized familiar voices affect the brain's neural responses. The activation of the prefrontal cortex and temporal cortex may reflect the multidimensional characteristics of voice familiarity processing, including emotion, memory, and cognitive functions.

At the same time, the innovation of this study lies in the adoption of a novel assessment method, which significantly differs from traditional speech synthesis evaluation approaches that mainly rely on subjective ratings (e.g., MOS (Mean Opinion Score), CMOS (Comparative Mean Opinion Score)) and audio analysis (e.g., MCD (Mel Cepstral Distortion), PESQ (Perceptual Evaluation of Speech Quality)). Traditional methods typically focus on the speech characteristics and auditory experience of the voice. In contrast, this study directly detects the brain's neural responses to AI-synthesized voices using fNIRS technology, providing more intuitive and vivid evidence. This approach not only reveals the impact of voice familiarity on brain activation but also captures the immediate neural responses generated by users during the listening process, offering a fresh research perspective for the application of AI voices in the emotional domain.

Additionally, this study emphasizes the potential applications of AI-synthesized voices in the realm of emotional support, suggesting that generating personalized familiar voices can effectively enhance emotional connection and user experience. In the field of mental health, using familiar voices synthesized by AI can alleviate feelings of loneliness and anxiety in patients, particularly for the elderly or those in long-term care. AI-generated voices can be personalized to simulate the voices of loved ones, providing emotional support to help patients cope with emotional challenges. This finding provides significant theoretical support for the application of AI technology in mental health and interpersonal relationship enhancement, showcasing the practical implications of this research.

In future research, it is recommended to further explore the emotional regulation role of AI-generated voices in specific contexts, particularly the responses under different familiarity levels. Although this study mainly focuses on the impact of AI-synthesized voices on brain responses, it is worth noting that familiar voices and emotionally relevant voices may overlap, especially in terms of their ability to trigger emotional reactions. Future studies could delve into the independent effects of AI-generated voices with different familiarity levels (e.g., familiar but non-emotional voices vs. emotionally relevant voices) on brain responses, in order to better understand the role of voice familiarity and emotional connection in brain activation. Additionally, exploring more application scenarios, such as virtual assistants, education, and entertainment, could fully unlock the potential of personalized voice synthesis technology. This would not only promote the development of the technology but also provide effective guidance for enhancing user experience.

In conclusion, this study innovatively applied fNIRS technology to directly detect the effects of AI voice synthesis through brain neural responses. The research shows that AI-synthesized familiar voices of loved ones can significantly influence the brain's neural responses and activate voice familiarity processing mechanisms. This finding provides important evidence for the application of personalized voice synthesis technology in the fields of emotion and cognition, helping to advance human-computer interaction and open a new chapter for richer, more humanized interactive experiences.

Data availability

Data have been provided in manuscripts or supplementary information documents. We provided transcripts of interviews with the participants. Please feel free to contact us if you need information such as informed consent form and assessment questionnaire.The data from this study are available upon reasonable request. For access to the data, please contact the designated author, Jiaju Li, at lijiaju0712@163.com.

Received: 4 November 2024; Accepted: 3 March 2025
Published online: 15 May 2025

References

1. Wang, Y. et al. Tacotron: Towards End-to-End speech synthesis. (2017). https://ui.adsabs.harvard.edu/abs/2017arXiv170310135W
2. Ren, Y. et al. FastSpeech: Fast, robust and controllable text to speech. (2019). https://ui.adsabs.harvard.edu/abs/2019arXiv190509263R
3. Huang, S. F., Lin, C. J., Liu, D. R., Chen, Y. C. & Lee, H. y. Meta-TTS: Meta-learning for few-shot speaker adaptive text-to-speech. *IEEE/ACM Trans. Audio Speech Lang. Process.* **30**, 1558–1571. https://doi.org/10.1109/TASLP.2022.3167258 (2022).
4. Hu, W. & Zhu, X. A real-time voice cloning system with multiple algorithms for speech quality improvement. *PLOS ONE* **18**, e0283440. https://doi.org/10.1371/journal.pone.0283440 (2023).
5. Dan, Q., Xukui, Y. & Honggang, Y. et al, Overview of recent progress in low-resource few-shot continuous speech recognition. *J. Zhengzhou Univ. (Engineering Science)* **44**, 1–9. https://doi.org/10.13705/j.issn.1671-6833.2023.04.014 (2023).
6. Efthymiou, F. & Hildebrand, C. Empathy by design: The influence of trembling AI voices on prosocial behavior. *IEEE Trans. Affect. Comput.* **15**, 1253–1263. https://doi.org/10.1109/TAFFC.2023.3332742 (2024).
7. Ahmed, S. & Chua, H. W. Perception and deception: Exploring individual responses to deepfakes across different modalities. *Heliyon* **9**, e20383. https://doi.org/10.1016/j.heliyon.2023.e20383 (2023).
8. Wong, N. et al. Voice assistants for mental health services: Designing dialogues with homebound older adults. *DIS. Designing Interact. Syst. (Conference)* **2024**, 844–858. https://doi.org/10.1145/3643834.3661536 (2024).
9. Abdollahi, H., Mollahosseini, A., Lane, J. T. & Mahoor, M. H. in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids).* 541–546.
10. Fogelson, D. M., Rutledge, C. & Zimbro, K. S. The impact of robotic companion pets on depression and loneliness for older adults with dementia during the COVID-19 pandemic. *J. Holist. Nurs.* **40**, 397–409. https://doi.org/10.1177/08980101211064605 (2021).
11. Stevenage, S. V. Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia* **116**, 162–178. https://doi.org/10.1016/j.neuropsychologia.2017.07.005 (2018).
12. Mathiak, K. et al. Who is telling what from where? A functional magnetic resonance imaging study. **18**, 405–409, (2007). https://doi.org/10.1097/WNR.0b013e328013cec4

13. Nakamura, K. et al. Neural substrates for recognition of familiar voices: A PET study. *Neuropsychologia* **39**, 1047–1054. https://doi.org/10.1016/s0028-3932(01)00037-9 (2001).
14. Schall, S., Kiebel, S. J., Maess, B. & Kriegstein, K. v. Voice identity recognition: Functional division of the right Sts and its behavioral relevance. **27**, 280–291, (2014). https://doi.org/10.1162/jocn_a_00707
15. Bitan, T. et al. Developmental changes in activation and effective connectivity in phonological processing. *NeuroImage* **38**, 564–575. https://doi.org/10.1016/j.neuroimage.2007.07.048 (2007).
16. Blank, H., Wieland, N. & von Kriegstein, K. Person recognition and the brain: Merging evidence from patients and healthy individuals. *Neurosci. Biobehav. Rev.* **47**, 717–734. https://doi.org/10.1016/j.neubiorev.2014.10.022 (2014).
17. Dixon, M. L., Thiruchselvam, R., Todd, R. & Christoff, K. Emotion and the prefrontal cortex: An integrative review. *Psychol. Bull.* **143**, 1033–1081. https://doi.org/10.1037/bul0000096 (2017).
18. Braunsdorf, M. et al. Does the Temporal cortex make Us human? A review of structural and functional diversity of the primate temporal lobe. *Neurosci. Biobehav. Rev.* **131**, 400–410. https://doi.org/10.1016/j.neubiorev.2021.08.032 (2021).
19. Naseer, N. & Hong, K. S. fNIRS-based brain-computer interfaces: A review. *Front. Hum. Neurosci.* **9** https://doi.org/10.3389/fnhum.2015.00003 (2015).
20. Wijeakumar, S., Huppert, T. J., Magnotta, V. A., Buss, A. T. & Spencer, J. P. Validating an image-based fNIRS approach with fMRI and a working memory task. *NeuroImage* **147**, 204–218. https://doi.org/10.1016/j.neuroimage.2016.12.007 (2017).
21. Zhang, Y. F., Lasfargues-Delannoy, A. & Berry, I. Adaptation of stimulation duration to enhance auditory response in fNIRS block design. *Hear. Res.* **424**, 108593. https://doi.org/10.1016/j.heares.2022.108593 (2022).
22. Zhou, Y., Chen, M., Lei, Y., Zhu, J. & Zhao, W. J. a. e.-p. VITS-based singing voice conversion system with DSPGAN post-processing for SVCC2023. (2023). https://ui.adsabs.harvard.edu/abs/2023arXiv231005118Z
23. Lester, C. et al. Comparing different motion correction approaches for resting-state functional connectivity analysis with functional near-infrared spectroscopy data. *11%J Neurophotonics*, 045001 (2024).
24. Zhang, N. et al. The effects of age on brain cortical activation and functional connectivity during video game-based finger-to-thumb opposition movement: A functional near-infrared spectroscopy study. *Neurosci. Lett.* **746**, 135668. https://doi.org/10.1016/j.neulet.2021.135668 (2021).
25. Li, Y., Song, F., Liu, Y., Wang, Y. & Ma, X. Relevance of emotional conflict and gender differences in the cognitive tasks of digital interface layouts using NIRS technology. *IEEE Access.* **9**, 17382–17391. https://doi.org/10.1109/ACCESS.2020.3048737 (2021).
26. Kamran, M. A., Mannan, M. M. N. & Jeong, M. Y. J. F. I. H. N. Cortical signal analysis and advances I. functional Near-Infrared spectroscopy signal: A review. 10 (2016).
27. Strangman, G., Culver, J. P., Thompson, J. H. & Boas, D. A. A quantitative comparison of simultaneous BOLD fMRI and NIRS recordings during functional brain activation. *NeuroImage* **17**, 719–731, (2002). https://doi.org/10.1006/nimg.2002.1227
28. Roswandowitz, C., Kathiresan, T., Pellegrino, E., Dellwo, V. & Frühholz, S. Cortical-striatal brain network distinguishes deepfake from real speaker identity. *Commun. Biol.* **7**, 711. https://doi.org/10.1038/s42003-024-06372-6 (2024).

Acknowledgments

We are deeply grateful to Shaomou Lu andQi Liu from the Integrated Circuit Innovation Class, Shaoxing University, for their invaluable technical support in data preprocessing andexperimental setup validation, which significantly contributed to the early stages of this research.

Author contributions

Z. and L. wrote the main manuscript text, L. completed the experimental part, and J. prepared the GPT-SoVITS project.Three authors have made equal contributions and wish to be listed as a joint work. All the authors reviewed the manuscript.

Declarations

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/s41598-025-92702-5.

Correspondence and requests for materials should be addressed to S.C.