Voice delivery enhances the therapeutic relationship between patients and AI therapy agents

George Prichard[1], Jessica McFadyen[1], Keno Juchems[1], Annamaria Balogh[1], Sashank Pisupati[1], Tobias U. Hauser[1,2,3], Ross Harper[1], & Max Rollwage[1],

[1] Limbic Limited, London, UK
[2] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London WC1B 5EH, UK
[3] Department of Psychiatry and Psychotherapy, Medical School and University Hospital, Eberhard Karls University of Tubingen, Tübingen, Germany
[4] German Center for Mental Health (DZPG), Tübingen, Germany

Abstract

Background: Mental health services face critical access barriers worldwide, with artificial intelligence (AI) emerging as a potential solution. While text-based AI therapy shows promise, the impact of voice-enabled AI therapy on therapeutic alliance remains unexplored.

Objective: To investigate whether voice-delivered AI therapy enhances therapeutic relationship compared to text-based delivery, and to examine associated impacts on user engagement and emotional support.

Methods: In a within-subjects design, 30 UK adults (mean age=41, 27% female) engaged with both voice and text-based AI therapy agents in counterbalanced order. Participants discussed personal problems in each modality. Outcomes included therapeutic alliance (WAI-SR), engagement metrics, and qualitative feedback.

Results: Voice interactions generated longer utterances (M=24 vs 12 words, $p<0.0005$) and stronger therapeutic alliance scores ($d=0.56$, $p=0.005$), with 66.7% of participants feeling a stronger therapeutic relationship with the voice agent. Voice interactions showed significantly higher emotional support ratings, though general usability and habit formation preferences were comparable between modalities.

Conclusions: Voice-enabled AI therapy significantly enhances therapeutic alliance and emotional connection compared to text-based delivery, suggesting important implications for improving mental health treatment accessibility and effectiveness.

Intro

Mental health services face overwhelming demand and a critical shortage of practitioners worldwide. In the UK, a recent survey showed only 21% of people felt they received adequate treatment in time (NHS Digital, 2023), while many patients with a mental illness receive no treatment at all (NIMH, 2022). This treatment gap has severe societal implications, with untreated mental health conditions contributing to reduced life expectancy, decreased productivity, and increased healthcare costs, estimated at $1 trillion annually in lost economic output (WHO, 2021).

The barriers to accessing mental health care are multifaceted: geographic limitations, cost constraints, stigma, and most critically, a severe shortage of qualified mental health professionals. Current estimates suggest that traditional training pipelines cannot fill the gap in the existing supply-demand imbalance (WHO, 2023).

Artificial Intelligence presents a promising solution to scale mental health support. Conversational AI, in particular, has demonstrated effectiveness in delivering evidence-based therapeutic interventions (Fitzpatrick et al., 2017). Companies like Limbic are pioneering Clinical AI developments, showing that AI-enabled therapy can provide accessible, cost-effective treatment options while maintaining clinical efficacy.

However, a fundamental challenge remains: the perception that emotional support and therapeutic relationship-building are uniquely human capabilities. This belief stems from our understanding of empathy as a complex interpersonal process involving subtle emotional cues, authentic emotional resonance, and genuine understanding (Rogers, 1957). Traditional therapeutic theory emphasizes that healing occurs through the therapeutic relationship itself, with meta-analyses suggesting that the quality of this relationship accounts for approximately 30% of therapeutic outcomes (Lambert & Barley, 2001).

Recent advances in voice-based AI interactions create new possibilities for bridging the perceived gap between human and machine therapeutic interactions. Voice communication carries rich emotional information through prosody, rhythm, and tone that text cannot convey (Scherer, 2003). These paralinguistic features have been fundamental to human emotional communication throughout our evolutionary history, suggesting that voice interaction might activate more primitive and authentic emotional processing pathways.

This study investigates the efficacy of therapeutic interactions delivered via generative voice or text AI on therapeutic connection and user experience. Voice-delivered therapeutic interactions offer multiple potential benefits: enhanced accessibility for visually impaired individuals or those with reading difficulties, reduced cognitive load compared to typing, and the ability to personalize the therapeutic experience through voice customization.

Our primary focus is the impact on the therapeutic relationship - the quality of rapport between patient and provider (in this case, the AI agent) - which consistently emerges as one of the strongest predictors of positive treatment outcomes across therapeutic modalities (Wampold, 2015). While technical AI capabilities have advanced rapidly, the ability to form

genuine emotional connections has been considered a uniquely human domain, making this an essential frontier for investigation.

We hypothesize that voice interactions, by more closely approximating natural human communication, can create stronger bonds and more effective emotional support between AI agents and users. This hypothesis challenges traditional assumptions about the exclusively human nature of therapeutic relationships and explores whether technological advances can bridge this perceived gap.

The implications of this research extend beyond immediate clinical applications. Understanding how voice interaction affects therapeutic alliance could inform the design of future AI-enabled mental health interventions and contribute to our broader understanding of human-AI emotional interaction. As mental health needs continue to outpace traditional care delivery systems, establishing effective methods for AI-delivered emotional support becomes increasingly crucial for global mental health.

Methods

Participants and Recruitment

The study recruited 30 participants through the Prolific research platform with specific pre-screening criteria to ensure participant safety and data quality. Eligible participants were UK residents aged 18-65, fluent in English, with no current severe mental health conditions. The final sample comprised 22 males and 8 females with a mean age of 41 years (SD=12.3).

Materials and Apparatus

Two distinct AI models were employed in this study. For text-based interactions, we utilized OpenAI's GPT-4 (gpt-4o-2024-08-06), while voice interactions were conducted using OpenAI's Realtime API (gpt-4o-realtime-preview-2024-10-01). Both models were configured with identical base prompts emphasizing person-centered counseling techniques and emotional support, with the voice model receiving an additional instruction to "speak quickly as if excited" to maintain natural conversation pace.

The primary outcome measure was the Working Alliance Inventory-Short Revised (WAI-SR; Hatcher & Gillaspy, 2006), which was adapted for the AI interaction context. This 12-item measure assesses three key dimensions of therapeutic alliance: Goals, Tasks, and Bond. Participant mood was assessed using a 7-point Likert scale ranging from "Very negative" (1) to "Very positive" (7), administered both pre- and post-interaction for each modality.

A comprehensive preference survey was developed to capture participants' experiences across multiple dimensions. The survey used a 5-point scale ranging from "Strongly Voice"

to "Strongly Text" to assess preferences in therapeutic relationship, usability, habit formation, and emotional support. Open-ended questions were included to gather qualitative feedback about participants' experiences with each modality.

Procedure

The study employed a within-subjects design with counterbalanced presentation of voice and text modalities to control for order effects. Participants first completed initial screening and provided informed consent. Before each interaction, participants completed the pre-interaction mood assessment. They then engaged in two therapeutic interactions (5-10 minutes each), one with each modality, discussing personal problems of their choice. The order of the agents was counterbalanced between participants. After each interaction, participants completed the post-interaction mood assessment and the WAI-SR. Following both interactions, participants completed the final preference survey and provided qualitative feedback about their experience.

Data Analysis

Engagement metrics (word count and message frequency) were compared between modalities using paired t-tests. Mood changes were analyzed using repeated measures ANOVA, with modality and time point as within-subjects factors. Preference distributions were analyzed using chi-square tests, and effect sizes were calculated using Cohen's d.

Qualitative data from participant feedback underwent thematic analysis following Braun and Clarke's (2006) approach. Two independent researchers coded the responses to ensure reliability, with disagreements resolved through discussion. The analysis focused on identifying recurring themes in participants' experiences with each modality and understanding the mechanisms underlying any observed preferences.

Results

Engagement Metrics and Interaction Patterns

Firstly, we were interested in the effects of the AI modality on general engagement with the AI agent. Over the whole conversation, users said significantly more words out loud than they typed per utterance (24 words vs 12 words, $t=4.94$, $P < 0.0005$). The total number of voice- or text-based messages, however, was not significantly different (see **Table 1**). This suggests that more efficient exchanges with the voice AI agent, leading to enhanced engagement.

	Average Words	Average Words	Average Messages	Average Messages
	text	voice	text	voice
assistant	34.19	41.00	12.40	12.25
user	12.00	24.56	11.40	11.40

Table 1: Average words and messages sent by each delivery method

Therapeutic Alliance and Emotional Support

Next, we investigated the impact of the voice modality on therapeutic relationship, emotional support. Participants reported stronger therapeutic connections during voice interactions compared to text ($t[29] = 3.07$, $p = 0.005$, cohen's $d = 0.56$; see Figure 1 and Table 2), with 66.7% of participants expressing a stronger therapeutic relationship with the voice AI agent than with the text based agent (see **Figure 1**). In line with these findings, participants also reported significantly higher emotional support through the voice than the text modality (see **Table 2**).

Mood Changes and General Usability

Changes in mood from pre- to post-conversation showed modest improvements across both modalities, but the difference between voice and text was not statistically significant ($p > .05$).

Regarding general usability and habit formation potential, preferences were more evenly distributed. No significant differences were found between modalities for ease of use ($p>.05$) or likelihood of regular use ($p>.05$). However, a trend toward voice preference was observed in both dimensions (usability: 53.3% voice preference; habit formation: 56.7% voice preference).

These results suggest a specifically pronounced effect on therapeutic relationship and emotional support through voice AI, above general effect on usability. This indicates that these novel capabilities might be especially transformative for building a strong therapeutic alliance between users and the AI therapist, as well as making users feel emotionally supported.

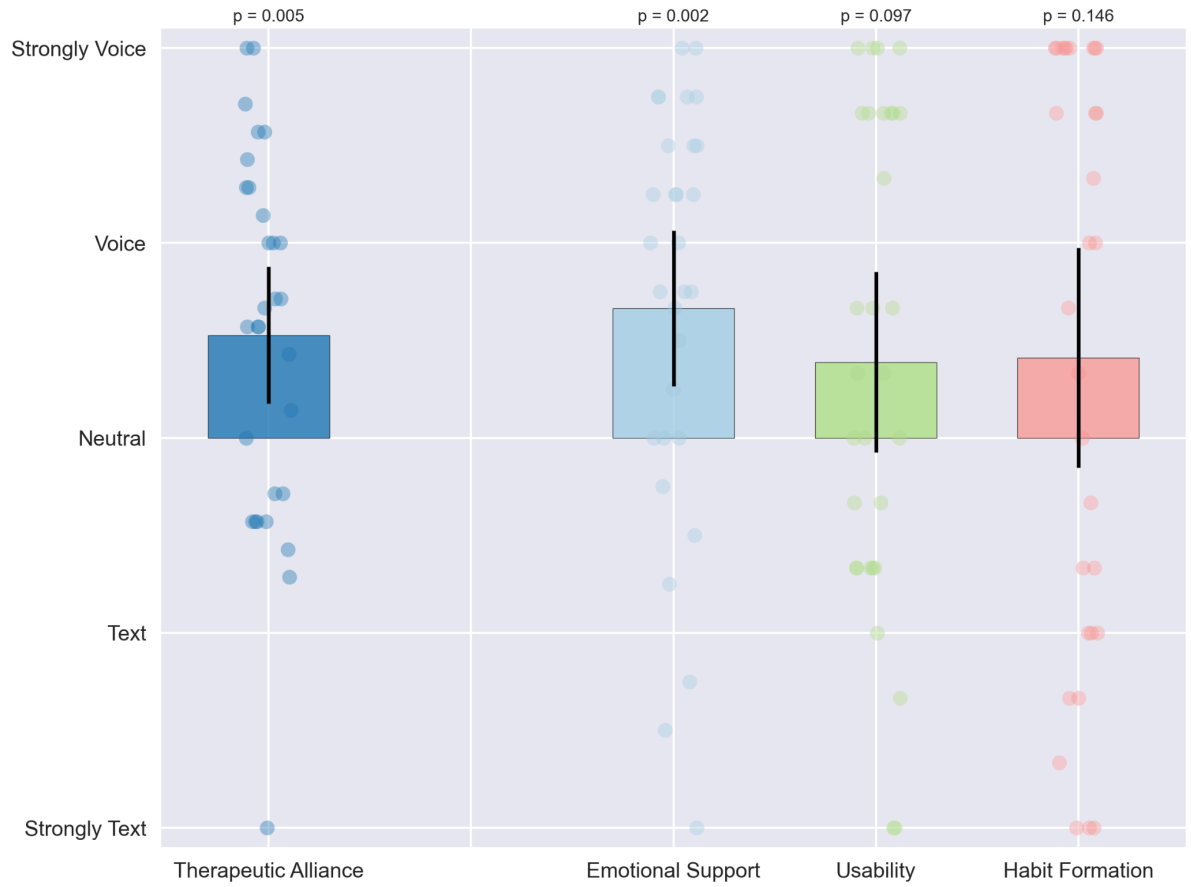

Figure 1: Preferences for voice vs text for each group of questions. P values are from single sample t-tests against a hypothesis of a mean preference of 0 (no preference for either voice or text). Error bars represent the 95% confidence intervals.

Group	p-value	t-statistic	Cohen's d	% Prefer Voice	% Prefer Text
Therapeutic Alliance	0.005	3.07	0.56	66.67	30.00
Emotional Support	0.002	3.40	0.62	70.00	20.00
Usability	0.097	1.72	0.31	53.33	36.67
Habit Formation	0.146	1.49	0.27	56.67	40.00

Table 2: Statistical test results comparing question groups. One-sample t-tests and Cohen's D effect sizes were calculated using a null hypothesis of 0 (neutral preference)."

Qualitative Analysis

Thematic analysis of participant feedback revealed three primary themes (see **Table 3** for sample quotes):

1. Enhanced Emotional Connection: Participants frequently described the voice interactions as more "natural" and "emotionally engaging." One participant noted, "Feels like a therapist somehow"

2. Modality-Specific Advantages: While voice was preferred for emotional expression, text interactions were valued for allowing more time to reflect and compose responses. A participant commented, "Voice felt more personal, but typing gave me time to think through my responses carefully."
3. Environmental Considerations: Several participants mentioned privacy and setting requirements as important factors in modality preference, particularly for voice interactions.

These qualitative insights provide context for the quantitative findings and highlight the nuanced nature of user preferences in AI-mediated therapeutic interactions.

	Sample Quotes
Text Chat	*I was pleased and surprised at how much of what I typed was taken in and responded to, very appropriately, by limbic. It was a very positive experience.*
Voice Chat	*I loved this. Feels like a therapist somehow.* *It made it easy for me to express my concerns without becoming emotional about it. I was able to express myself and, at the same time, retain composure.* *Unbelievable how clever that was* *This was very good. The voice sounded quite natural and there was very little delay in limbic constructive responses.*
Comparison	*They were similar in the quality of their responses, however the voice felt more emotionally stimulating and thought provoking* *I was surprised by the effectiveness of the voice interactions, my only concern is the potential privacy issues. Other people may overhear personal and private information so the suitable environments for the voice interactions are less numerous.* *Totally mindblown by the voice chat but i think both were really effective*

Table 3: Sample qualitative feedback from users after either the voice or text modality, or after the comparison survey.

Discussion

The present study provides compelling evidence that voice-enabled AI therapy can enhance the therapeutic relationship between users and AI agents compared to traditional text-based interactions. Our findings demonstrate significant advantages of voice delivery, particularly in fostering emotional connection and therapeutic alliance, while also revealing important nuances in user preferences and practical considerations.

Therapeutic Alliance and Emotional Connection

The substantial effect size (d=0.56) observed in therapeutic alliance scores represents a meaningful improvement in the quality of therapeutic relationships through voice interaction. This finding is particularly noteworthy given that the interactions were brief (5 minutes) and single-session, suggesting that voice modality can rapidly facilitate stronger therapeutic bonds.

Qualitative participant feedback explicitly stated that the AI voice was more “emotionally stimulating” or “made it easy for me to express my concerns”, which aligns with our quantitative results. Taken together, these results point towards a specific mechanism through which voice capabilities could have unique benefits for AI therapeutic interactions.

The enhanced emotional support reported in voice interactions likely stems from several mechanisms. First, the prosodic elements of voice communication may enable more nuanced emotional expression and recognition (Scherer, 2003). Second, the increased verbal output in voice interactions suggests greater engagement and disclosure, which are crucial components of therapeutic processes (Pennebaker, 1997).

Practical Implications and Implementation Considerations

While our results strongly favor voice delivery for therapeutic alliance, the findings regarding usability and habit formation were more nuanced. The absence of significant differences in these domains suggests that the advantages might be especially focussed on the positive emotional effects of voice interaction for therapy settings. However, environmental constraints, such as privacy requirements and ambient noise, emerge as important factors in implementation decisions.

The finding that 30% of participants preferred text-based interaction reinforces the importance of offering modality choice in therapeutic AI applications. This preference heterogeneity suggests that optimal implementation might involve flexible, multi-modal delivery systems that allow users to switch between voice and text based on their circumstances and preferences.

Limitations and Future Directions

Several limitations warrant consideration. First, the single-session design, while demonstrating immediate effects, cannot speak to the longitudinal impact of voice versus text delivery. Future research should examine how these effects evolve over extended therapeutic relationships and multiple sessions.

Second, our sample's demographic composition (predominantly male, UK-based) may limit generalizability. Cultural differences in verbal communication styles and therapeutic expectations could influence the relative benefits of voice interaction. Future studies should investigate these effects across diverse populations and cultural contexts.

Third, while we controlled for content by using identical base prompts, the addition of the "speak quickly" instruction in the voice condition introduces a potential confound. Future research should more precisely match interaction parameters across modalities.

Important directions for future research include:

1. Longitudinal studies examining the stability and development of therapeutic alliance over time

2. Investigation of hybrid approaches combining voice and text modalities
3. Exploration of voice customization effects on therapeutic outcomes
4. Assessment of treatment efficacy outcomes beyond alliance measures

Clinical and Technological Implications

These findings have significant implications for the development and deployment of therapeutic AI systems. The marked improvement in therapeutic alliance through voice delivery suggests that voice capability should be a priority feature in future therapeutic AI applications. However, developers must carefully consider the technical challenges of implementing voice interactions, including latency management, privacy protection, and output safety checking.

Conclusion

This study demonstrates that voice delivery significantly enhances therapeutic interactions with AI agents. Most notably, voice interfaces strengthen the therapeutic relationship - a cornerstone of effective therapy that was previously thought to be uniquely human. Given that AI for mental healthcare has the capacity to scale to benefit people not receiving adequate support, optimizing the delivery method has the potential to improve wellbeing for millions of people.

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Mental Health, 4(2), e19.

Hatcher, R. L., & Gillaspy, J. A. (2006). Development and validation of a revised short version of the Working Alliance Inventory. Psychotherapy Research, 16(1), 12-25.

Horvath, A. O., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. Psychotherapy, 48(1), 9-16.

Lambert, M. J., & Barley, D. E. (2001). Research summary on the therapeutic relationship and psychotherapy outcome. Psychotherapy: Theory, Research, Practice, Training, 38(4), 357-361.

NHS Digital. (2023). Mental Health Services Monthly Statistics. Retrieved from https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-services-monthly-statistics

NIMH (National Institute of Mental Health). (2022). Mental Health Information Statistics. Retrieved from https://www.nimh.nih.gov/health/statistics/mental-illness

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. Psychological Science, 8(3), 162-166.

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. Journal of Consulting Psychology, 21(2), 95-103.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech Communication, 40(1-2), 227-256.

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. World Psychiatry, 14(3), 270-277.

WHO (World Health Organization). (2021). Mental Health Atlas 2020. World Health Organization.

WHO (World Health Organization). (2023). Mental Health Workforce Statistics. World Health Organization.