La sala hace la mitad del trabajo

Cuerpo, público y conducta de un agente de voz en tres eventos

Cris Olmedo y Kurt Sommer
Grupo de investigación Divina Machina
Documento de trabajo · Estudio observacional
DM-2026-06, versión 1.0 · 30 de octubre de 2026

Resumen. Un personaje de voz es tres cosas a la vez: un modelo, un cuerpo y la gente que tiene delante. Comparamos el mismo personaje, GAIA 2147, en una cabina telefónica con luz, pantalla e impresora, en tres eventos: Sónar+D (701 llamadas), IFA (274) y Barcelona Design Week (543).

La sala marcó las llamadas más que el guion. En IFA, el 49 % (IC 95 % 43–55 %) de quienes descolgaron no dijo nada, frente al 22 % en Barcelona. En Sónar+D, el 36 % (IC 95 % 33–40 %) de los turnos del visitante eran la propia voz del agente recogida por el micrófono. El modelo apenas usó su cuerpo: en Barcelona, solo el 5,2 % (IC 95 % 3,6–7,4 %) de las llamadas incluyó una herramienta de luz, pantalla o papel. La gente fue amable con la máquina, la puso a prueba y le habló sobre todo de máquinas y de futuro.

Proponemos repartir responsabilidades: el modelo decide el sentido; la sala, en el dispositivo, produce la presencia. Cada evento es una única muestra de sala, público y configuración, y lo tratamos así.

Palabras clave: agentes encarnados; espacios públicos; herramientas; conducta del usuario; eco acústico; instalaciones interactivas.

1. Introducción

Casi toda la atención de diseño de un agente de voz va al modelo. Nuestros registros dicen que el cuerpo y el público decidieron tanto como él. GAIA 2147 contesta en una cabina de Telefónica con un auricular, una pantalla de 20 por 4 caracteres, una tira de 60 luces y una impresora térmica. El modelo puede usar las cuatro como herramientas; el prompt le pide hacerlo poco [1]. Siempre trabajamos con ElevenLabs y Vidiv; estos datos son del agente de ElevenLabs.

Este trabajo pone uno junto a otro tres hallazgos: la sala cambia las llamadas, el modelo casi no usa su cuerpo y la gente hace cosas que una transcripción apenas recoge.

2. Datos y método

Los tres eventos difieren en público, idioma, ruido, pila de voz, prompt, saludo y límite de tiempo. Por eso no ordenamos causas: describimos cada sala con sus intervalos y medimos la heterogeneidad entre eventos con chi cuadrado y la V de Cramér, con corrección de Holm cuando hay varias conductas [2]. Proporciones con Wilson [3]; medianas con bootstrap; tasas por turno con bootstrap de conversaciones [4]. Las conductas, los idiomas y los temas se detectan con léxico en siete idiomas: cada proporción es un mínimo [DEF]. El eco marca un turno del visitante de cuatro o más palabras con al menos el 60 % en el turno anterior de GAIA:

$$\mathrm{eco}(v_i) = \mathbb{1}\Big[|W(v_i)| \geq 4 \ \wedge \ \tfrac{|W(v_i) \cap W(g_{i-1})|}{|W(v_i)|} \geq 0{,}6\Big]. \quad (1)$$

3. Tres salas

La tabla 1 resume las salas [MED]. En IFA casi la mitad de las llamadas duró menos de diez segundos y otra mitad fue muda. Las diferencias entre eventos son claras en llamadas cortas (V = 0,22) y mudas (V = 0,20), ambas con $p < 0{,}001$. IFA juntó la pila más lenta, un público de feria que va de stand en stand y que hablaba sobre todo inglés y alemán; no sabemos cuál pesó más [INT].

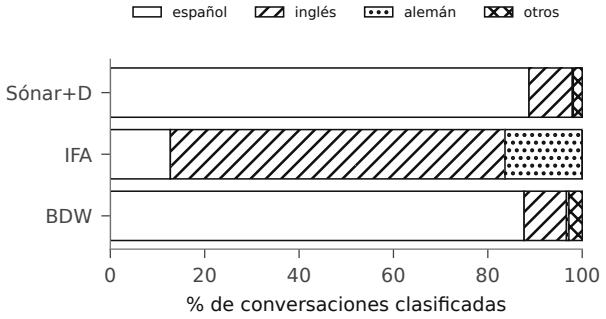


Figura 1. Idioma del visitante por evento, entre conversaciones clasificadas.

Quién estaba al otro lado. Cada sala tuvo su idioma (figura 1): el 89 % de las conversaciones clasificadas fue en español en Sónar+D; en IFA, el 71 % en inglés y el 16 % en alemán; en Barcelona, el 88 % en español. GAIA cambió de idioma a mitad de llamada en el 11,7 %, 12,8 % y 6,4 % de las conversaciones.

La sala contesta. En Sónar+D el 36 % (IC 95 % 33–40 %) de los turnos del visitante eran eco, frente al 4 % en IFA y el 10 % en Barcelona. El eco infla lo que parece decir el visitante (16,4 palabras por turno en Sónar+D frente a 7,5 en Barcelona) y hace que el agente se conteste a sí mismo. En Barcelona lo combatimos con una puerta de ruido y bajando el micrófono mientras habla GAIA [MED]. No documentamos el altavoz ni el volumen de Sónar+D, así que no podemos atribuir la causa allí.

Cuando el público pone a prueba a la máquina. En Sónar+D, el 9,7 % (IC 95 % 7,7–12,1 %) de las llamadas lo cortó un guardarraíl de seguridad de la plataforma: 53 por contenido sexual, 7 por autolesión, 4 por lenguaje ofensivo y 3 por violencia [MED]. En IFA no hubo

Tabla 1. Tres salas. Resultados por evento.

Evento	n	Mediana (s)	Corta	Muda	Límite	Seguridad	Anómala	Eco
Sónar+D	701	58 [47; 66]	27 % [24; 30]	28 % [25; 31]	11 %	9,7 %	7,4 %	36 %
IFA	274	14 [9; 18]	46 % [40; 52]	49 % [43; 55]	3 %	0,0 %	1,1 %	4 %
BDW	543	80 [71; 88]	17 % [14; 21]	22 % [19; 26]	33 %	0,9 %	0,7 %	10 %

Corchetes, IC 95 % (bootstrap para la mediana, Wilson para proporciones). Eco, turnos del visitante que repiten el turno anterior de GAIA.

Tabla 2. Llamadas a herramientas decididas por el modelo, por evento.

Herramienta	Sónar+D	IFA	BDW
Luz	0	2	7
Pantalla	0	0	2
Impresión	0	1	17
Ticket	0	0	4
Cambio de idioma	89	40	39
Colgar	29	3	16

Número de llamadas a cada herramienta decididas por el modelo. Lo que la cabina hace por su cuenta no queda registrado aquí.

ninguno. Los registros no dicen si lo activó el visitante o el agente.

4. El cuerpo que el modelo no usa

En Barcelona el modelo usó al menos una herramienta de cuerpo en el 5,2 % (IC 95 % 3,6–7,4 %) de las llamadas; en las de un minuto o más, el 8,7 % (IC 95 % 6,1–12,3 %). En IFA, el 1,1 %; en Sónar+D, ninguna (tabla 2). Las herramientas más usadas fueron las de sistema: cambiar de idioma y colgar. Entre las de cuerpo ganó el papel. En el conjunto, 31 resultados de herramienta volvieron con error [MED].

Poco fue la instrucción y poco fue lo que obtuvimos: pedimos a un modelo de lenguaje que hiciera de iluminador como tarea secundaria, en mitad de una conversación, con una regla que le pedía contenerse [INT]. Lo que hizo parecer viva a la cabina lo decidía la cabina: una luz que sigue el ánimo de cada frase del visitante con reglas locales, una pantalla que murmura, el tono de la línea, el papel que sale al colgar. Nada de eso queda en los registros de conversación, y es lo primero que añadiríamos a la instrumentación.

La literatura sobre agentes con cuerpo apunta en la misma dirección: en una revisión de 33 experimentos, los robots que compartían espacio con la persona se valoraron mejor que los mismos agentes en pantalla, y pesó más la presencia física que el cuerpo en sí [5].

5. Lo que hizo la gente

Entre las 1067 conversaciones en que el visitante habló, la tabla 3 recoge seis conductas. El 28 % (IC 95 % 26–31 %) empezó con un saludo propio, la primera señal de que acepta la ficción de una llamada. El 5 % (IC 95 % 4–6 %) preguntó si GAIA era real, un robot o quién era; esas llamadas duraron 125 s de mediana frente a 89 s, lo que leemos como señal de interés y no como causa, porque una llamada más larga deja más sitio para la pregunta [INT].

La gente es amable con máquinas que no lo necesitan [6]. El 8 % (IC 95 % 7–10 %) dio las gracias, más en Sónar+D (11 %) que en IFA (4 %); es la única conducta que difiere entre eventos tras la corrección ($p = 0{,}01$). La risa casi no llega a la transcripción (1 %): es la conducta más visible en la cabina y la menos visible en los datos.

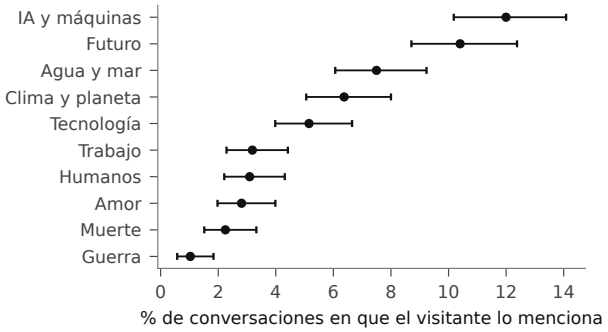


Figura 2. Temas mencionados por el visitante, con IC 95 % de Wilson. Detección por léxico; son mínimos.

Respuestas cortas a una máquina curiosa. El 24 % de los turnos del visitante tenía una o dos palabras, más en Barcelona (32 %) que en Sónar+D (15 %). Barcelona fue también donde más preguntó el agente (DM-2026-05). Cuanto más entrevista la máquina, más corto responde la gente [INT].

De qué quería hablar la gente. En el 35 % de las conversaciones el visitante tocó alguno de los diez temas que seguimos (figura 2). Primero las máquinas (12 % (IC 95 % 10–14 %)) y el futuro (10 %); el amor (3 %) y la muerte (2 %) fueron raros. El clasificador de sentimiento del proveedor, no validado por nosotros, etiquetó el 66 % de 562 llamadas como neutras.

6. Propuesta: el sentido para el modelo, la presencia para la sala

El modelo decide el sentido: prometer un papel, una palabra que se lee, un cambio de luz cuando la conversación gira. La sala produce la presencia: barata, inmediata y siempre activa, en el dispositivo, sin esperar al modelo [HIP]. Una señal de turno en la mano importa sobre todo a los niños, que aprenden cuándo ha terminado de hablar una máquina [7].

7. Amenazas a la validez

- Cada evento es una muestra única de sala, público y configuración; las diferencias son asociaciones.
- Detección léxica: conductas, idiomas y temas son mínimos; el eco y el reconocimiento emborronan el lado del visitante.
- Solo se registra lo que decide el modelo; lo que hace la cabina por su cuenta no tiene recuentos.

Tabla 3. Conductas del visitante por evento, con heterogeneidad entre eventos (V de Cramér) y p ajustada por Holm.

Conducta	Total [IC 95 %]	Sónar+D	IFA	BDW	V	p Holm
Saluda	28 % [26; 31]	25 %	24 %	33 %	0,09	0,09
Pregunta si es real	5 % [4; 6]	6 %	1 %	5 %	0,07	0,26
Da las gracias	8 % [7; 10]	11 %	4 %	6 %	0,11	0,01
Se despide	6 % [5; 7]	7 %	6 %	5 %	0,05	0,85
Ríe (en texto)	1 % [0; 1]	1 %	1 %	0 %	0,04	0,85
Dice palabrotas	2 % [2; 3]	3 %	1 %	2 %	0,05	0,85

Entre conversaciones en que el visitante habló. Detección por léxico en siete idiomas; cada proporción es un mínimo.

Tabla 4. Kit de atmósfera. Las filas propuestas son hipótesis.

Elemento	Quién lo activa	Estado
Luz	La sala por ánimo y turnos; el modelo en los giros	Construido
Pantalla	El modelo, palabras para leer; la sala, murmullos	Construido
Papel	El modelo lo promete; la cabina imprime al colgar	Construido
Sonido de línea	La sala, en cada llamada	Construido
Ambiente sonoro	La sala, según la tensión	Propuesto [HIP]
Vibración de turno	La sala, en el auricular	Propuesto [HIP]

- Límites de tiempo distintos por evento truncan la duración de forma distinta.

8. Recomendaciones para la comunidad

1. Reportar el eco, las llamadas mudas y los cortes de seguridad por sala: hablan de la sala antes que del personaje.
2. Registrar cada acción del entorno con su hora junto al identificador de la conversación.
3. No delegar la presencia en el modelo: lo ambiental, en el dispositivo y al instante.
4. Añadir una hoja de observación por hora; lo que no llega a la transcripción, como la risa o los grupos, solo lo ve una persona.

9. Datos y código

Dataset gaia-2147 [1]; cifras de `analysis/analyze.py`. No se guardó texto de los visitantes.

10. Conclusiones

La sala decidió quién descolgaba, en qué idioma, con cuánto eco y con qué ganas de jugar. El modelo, libre de usar su cuerpo, casi no lo hizo. La presencia de un personaje de voz se construye fuera del modelo, en la sala y en el dispositivo.

Firman. Cris Olmedo y Kurt Sommer, Grupo de investigación Divina Machina. *Cómo citar*: Olmedo, C. y Sommer, K. (2026). La sala hace la mitad del trabajo. Documento de trabajo DM-2026-06, v 1.0. Grupo de investigación Divina Machina. Divina Machina by Qs Ventures.

Referencias

Fuentes consultadas hasta el 29 de octubre de 2026.

[1] C. Olmedo and K. Sommer. GAIA-2147. descriptor de datos de 1527 conversaciones con un agente de voz en espacios públicos. Documento de trabajo DM-2026-09. Grupo de investigación Divina Machina, 2026.

[2] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[3] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

[4] C. A. Field and A. H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society. Series B*, 69(3):369–390, 2007. doi: 10.1111/j.1467-9868.2007.00593.x.

[5] J. Li. The benefit of being physically present. a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77:23–37, 2015. URL https://napier-repository.worktribe.com/output/3636900.

[6] C. Nass, J. Steuer, and E. R. Tauber. Computers are social actors. In *CHI '94*, pages 72–78, 1994. doi: 10.1145/191666.191703.

[7] J. Festerling and I. Siraj. Anthropomorphizing technology. a conceptual review of anthropomorphism research and how it relates to children's engagements with digital voice assistants. *Integrative Psychological and Behavioral Science*, 56:709–738, 2022. URL https://pubmed.ncbi.nlm.nih.gov/34811705/.