# The art of conversation

## From the *Before* trilogy to the design of conversational agents

Cris Olmedo and Kurt Sommer
Divina Machina Research Group
Conceptual article and research proposal
DM-2026-01, version 0.2 · 29 October 2026

**Abstract.** A conversation can change the meaning of an encounter, a memory or a decision. This article starts from an interpretive reading of the *Before* trilogy to investigate how time, attention, space and memory take part in that transformation, and what questions they raise for the design of conversational agents.

Through a targeted review of interaction research, film interviews and technical documentation, we propose to distinguish the availability of a response from the moment at which it is appropriate to intervene. We develop measurement definitions, decision models and an architecture to coordinate voice, shared objects and memory. The design hypothesis is that this coordination can sustain continuity and understanding without reducing the person's control.

The contribution is conceptual and methodological. Published figures keep their statistic, unit and measurement conditions; research results, vendor statements and evaluator reports are kept distinct. The equations make operational decisions and their assumptions explicit; they are not models fitted to our own data. No systematic temporal annotation of the films and no experimental evaluation of the proposed systems has been carried out. The article presents a testing programme, with measurement, analysis and preregistration requirements still to be executed.

**Keywords:** conversation; interaction design; voice agents; temporality; narrative; memory; multimodal interaction.

## 1. Introduction

### 1.1. Problem and working definitions

In our reading of *Before*, a conversation makes a journey meaningful because those who travel it come to understand something different by the end. Words take part in that change together with glances, places and the time available. We take that reading as the starting point for a research question: which coordination decisions allow an exchange to widen what a person can understand, express or decide? The proposal seeks to describe observable, testable conditions, while keeping the difference between a scene and the experience of someone conversing with a system.

In this article we call conversational narrative a sequence of exchanges whose development modifies a shared situation and allows continuity to be recognised between what came before and what follows. That continuity does not require a closed plot. It may consist of finding a better question, acknowledging a disagreement or deciding that we do not yet want to decide. An excellent administrative conversation can end quickly; a creative exploration can benefit from a digression.

We call conversational atmosphere the perceived quality of being in a situation with a certain rhythm, room for action, attention and continuity. It is a working definition, not a physical magnitude or an automatic reading of emotional state. Atmosphere is studied through observable behaviour, interviews and participants' ratings. The system can record a pause; it cannot conclude from that datum alone that there is trust, sadness or intimacy.

*A pause has a duration we can measure and a meaning we must ask about of whoever lived it.* This formulation guides the research: the temporal record and the account of experience provide different information and must keep their provenance.

The central hypothesis is that designing that atmosphere requires coordinating language, temporality, space and memory. The language model generates or interprets content; the product establishes who may intervene, which objects appear and what persists. If those decisions are left to the chance of infrastructure, the experience is also left to that chance.

### 1.2. Scope and approach of the article

The review is narrative, targeted and exploratory; it is not a systematic review or a meta-analysis. Research articles, primary interviews, vendor documentation and reports from the evaluator who ran each benchmark are prioritised. Sources are selected for their relation to turn coordination, shared understanding and narrative design. No exhaustive search, screening and exclusion flow has been reconstructed; therefore, no complete coverage or absence of selection bias is claimed.

Product information has a cut-off date of 29 October 2026. The films are discussed with spoilers. Numbered references are at the end; scenes are identified by their situation, without inventing timestamps that could vary between editions.

The film analysis presented is interpretive and relies on the cited sources. Systematic annotation of scenes and comparison between coders are part of the proposed research programme.

### 1.3. Traceability and status of claims

We distinguish published empirical results, vendor technical statements, third-party evaluations, mathematical definitions, film readings and design hypotheses. Each category licenses different inferences. A commercial figure does not become our own measurement by being cited; an observed association does not by itself

establish a causal relation; a correctly written equation does not show that its parameters have been estimated.

For the central figures we record source, consultation date, variable, unit, statistic, population or system and known conditions. We keep apart *not reported in the source*, *not verified in this review* and *not measured by this project*. The provenance and limits of the figures are collected in the supplementary material. Poetic sensitivity enters the questions and the interpretation of scenes; measurements keep their definitions and their uncertainty.

## 2. Film analysis of the *Before* trilogy

### 2.1. Shared authorship and the construction of naturalness

*Before Sunrise* was released in 1995, *Before Sunset* in 2004 and *Before Midnight* in 2013. The credited screenplay of the first is by Richard Linklater and Kim Krizan. The two later films credit the screenplay to Linklater, Julie Delpy and Ethan Hawke; Sunset also credits the story to Linklater and Krizan. This authorship should be preserved: attributing all of the trilogy's language to a single director would erase part of its method. [1–3]

Perceived spontaneity does not show improvisation on set. In an interview about Midnight, Linklater describes a meticulous construction and Delpy explains that they rehearsed for two weeks with an essentially finished script. In his recollection of Sunrise, Linklater describes an intense rewrite together with the actors. Each one's contribution ran through both characters. [4, 5]

*Film reading.* We understand that prior work as a condition of the naturalness we perceive in the relationship. For an agent, the useful translation is to prepare conditions, limits and repair capacities. It does not consist of forcing a person to play a role in a foreseen story.

### 2.2. *Before Sunrise*: discovery and shared attention

Two strangers decide to extend an encounter and walk through Vienna until morning. In our reading, limited time turns each detour into a choice; the conversation tests affinities, differences and forms of exposure, while the city offers situations that redirect the dialogue. In a record shop's listening booth, attention to a song makes glances and hesitations visible. [1, 5] In the café they act out imaginary calls to their friends. [6]

Our reading is that narrative interest arises from a gradual transformation of permission: what can I ask you now, what can I acknowledge in front of you, what can I risk without the encounter breaking. Intellectual content matters, but it also works as an indirect way of feeling out the relationship. An idea about life can be, at once, an offer of closeness or a protection against it.

We read the silence in the booth as a form of attention sustained by a shared activity. Both are listening to something; they are not simply waiting for someone to generate a sentence. Linklater recounts that he saved listening to the Kath Bloom recording for the shoot, even though the scene was rehearsed. This allows a prepared structure to be separated from a reaction that keeps its freshness. [5]

*Design proposal.* An agent can offer a shared object, an image or a piece of sound, and let the person choose what to do with it. It should be able to remain available without producing continuous commentary. The digital equivalent of the walk would be an exploration with possibilities of detour and return, not an itinerary that simulates freedom while leading to a predetermined conversion.

### 2.3. *Before Sunset*: reunion and reinterpretation of the past

The reunion in Paris happens nine years later. A travel commitment limits the time available. The route from the bookshop to other spaces, including Céline's apartment, accompanies the revelation of different versions of what happened and of what their lives have become. [2, 7] We read the song and the final gesture as the suggestion of a decision whose reach remains open, without turning it into an exhaustive explanation.

Linklater explains that they organised an emotional progression: the café allows a first, still protected honesty; the car concentrates greater exposure. Running out of time favours that progression. He also recounts that an imitation of Nina Simone by Delpy during the writing let him recognise the ending they were looking for. [7]

Our reading places memory as active material. A memory is not valuable only because it is retrieved: what matters is what it means now to each person. The past returns with disagreement, omission and reinterpretation. Conversation becomes story when it forces the revision of a version that seemed stable.

*Design proposal.* An agent's memory should be able to present itself with provenance and openness to correction: "Last time we noted this; does it still represent you?". Retrieving a datum does not license treating it as permanent identity. Continuity requires allowing someone to change, contradict a preference or withdraw an earlier interpretation.

### 2.4. *Before Midnight*: living together and repair

In Greece, the couple shares a life crossed by responsibilities, children and expectations. The conversations in the car, the meal with other people and the hotel shift attention from initial discovery to living together and its conflicts. [3, 4, 8] We read the final encounter as an attempt at repair that keeps the tensions exposed.

Our reading is that shared memory can widen the meaning of a sentence until it becomes hard to bear. A practical remark carries earlier conversations with it. Verbal precision does not guarantee understanding; knowing the other also makes it possible to anticipate them unfairly. The story can move forward through a reformulation, a withdrawal or an offer that does not demand immediate capitulation.

*Design proposal.* An agent's quality should be evaluated when it is wrong, when the user corrects it or

**Table 1.** Resources of the trilogy and proposed translations to conversational design.

| Resource | Function in our reading | Design translation | Limit |
|---|---|---|---|
| Available time | Gives weight to continuing or saying goodbye | Agree on duration and allow closing | Do not manufacture urgency |
| Route | Changes the frame and opens topics | Navigate objects or perspectives | Do not distract by default |
| Subtext | Allows indirect meanings | Ask for clarification tactfully | Do not infer hidden wishes as facts |
| Shared object | Distributes attention | Manipulable image or document | Do not impose a reading |
| Recurrence | Links separate moments | Retrieve a reference with provenance | Do not freeze identity |
| Repair | Changes what can happen next | Admit error and revise context | Do not demand reconciliation |

*Note.* Film reading and design proposal of the article; not empirically validated equivalences.

when an inference about their life proves inappropriate. Repairing means acknowledging the specific error, updating the context and handing back control. Accumulating data about a person does not grant a licence to interpret their motivations or to insist on a supposed emotional truth.

### 2.5. Scales of temporality

The trilogy articulates the instant of the exchange, the duration of an encounter and the years between encounters. Sunrise condenses several hours; Sunset is conceived in real time. That does not mean it is shot in a single take. Linklater relates his interest in time to Tarkovsky and points to Rohmer's precision, including *The Green Ray* as an influence. These statements support a specific genealogy, without turning every aesthetic similarity into a demonstrated influence. [9]

We propose to distinguish a micro scale, where voice onsets, pauses and overlaps matter; a scene scale, where an activity or question keeps continuity for several minutes; and a trajectory scale, where objects and agreements reappear across sessions. Each scale requires different tools. Improving time to first audio does not fix a misattributed memory, and retrieving a memory does not fix constant interruption.

### 2.6. Transferable resources for design

Table 1 summarises the proposed design correspondences and their limits.

## 3. Background and foundations of conversational interaction

### 3.1. Film tradition and the limits of transfer

Cinema offers a laboratory of forms, but its scenes are constructions for spectators. Editing can hide friction, framing can make a glance legible and an actor can sustain an ambiguity that a voice agent has no means to represent. That is why we should not read a film as a measurement of spontaneous conversation.

Rohmer allows us to study how speech organises the distance between a justification and a choice. *My Dinner with André*, by Louis Malle, written by André Gregory and Wallace Shawn, allows us to observe how a contained spatial situation can sustain deep differences of experience and perspective. The process Gregory describes includes recorded conversations, selection and rehearsal. As a later counterpoint, Ryusuke Hamaguchi explains a practice of repeated reading and attention to voice and body that makes it possible to investigate listening beyond verbal meaning. These are analytical comparisons; we do not claim a causal chain between all the works. [9–11]

*After Yang* raises an especially fertile question for systems with memory. Kogonada explains his search for a memory interface that did not look like a desktop with folders, and formally distinguishes Yang's recorded memories from reinterpreted human memories. The decision is cinematic, not a scientific description of how the brain remembers. Our reading is that the archive can take part in atmosphere: the way of returning to something also changes its meaning. [12]

The difference that should guide any transfer is agency. In a film, the characters serve a work. In a conversation with a system, the person must be able to leave, disagree, lower the intensity, ask for facts or declare that they do not want to remember. Narrative design has to support those decisions. A tool does not fail because the user changes the expected story.

This rules out using conflict, scarcity of time or intimacy as tricks to prolong a session. A system can sustain an atmosphere without attributing human experiences to itself or claiming emotional reciprocity. The proposed aim is understandable, free participation, compatible with the person's purposes.

### 3.2. Organisation of turns

The work of Sacks, Schegloff and Jefferson describes the local organisation of turns and its adaptation to participants. It provides a basis for understanding conversation as a coordinated activity. It does not establish a universal optimal interval. [13]

Stivers and colleagues analysed functional yes/no questions and their answers in ten languages, on a base of 101 conversations. They reported a pooled mean of 208 ms, a median of 100 ms and a mode of 0 ms. The onset of the answer could be vocal or gestural; instrumental measurements were made in 10 ms increments

and rounded to 100 ms. Therefore, the zero mode does not imply exact synchronisation and the variable does not necessarily coincide with an agent's audible onset. These statistics describe that corpus; they do not set a universal optimal interval. [14]

Levinson and Torreira review how brief conversational gaps coexist with language production times of 600 ms or more for certain tasks. The explanation involves anticipation and overlap between comprehension and preparation. The 600 ms are not a constant of human thought. [15]

*Design proposal.* The agent should be able to prepare hypotheses while listening and discard them if the user continues. This anticipation can reduce waiting, but it must not execute external actions or play a conclusion before having enough context. Understanding incrementally and acting prematurely are different decisions.

### 3.3. Contextual effects of speed and pause

The conversation studies of Templeton and colleagues related response gaps to ratings of connection collected while reviewing recordings. In a separate experiment, the gaps of recorded fragments were modified and the connection perceived by external listeners was assessed. The causal inference of that manipulation concerns third-party perception; the studies of lived conversation are observational. The result does not show that speeding up an agent produces greater connection for its user. [16]

Another work by the same group observed different associations between gaps longer than two seconds and declared connection among friends and strangers. The threshold was an analytical definition; friendship and pause duration were not experimentally assigned. During those gaps laughter and other non-verbal behaviour could be observed. Their interpretation as shared reflection is a hypothesis of the authors, not a direct reading of an internal state. [17]

The transfer hypothesis is therefore contextual. An immediate answer to a closed question can be appropriate. The same entry during a word search can feel invasive. A silence accompanied by an image that both contemplate does not necessarily have the same function as a wait without any signal while a tool fails.

### 3.4. Shared understanding and situated action

Clark and Brennan describe grounding as the construction of evidence of understanding sufficient for the current purpose. Goodwin studies how speech, body and materials of the environment are integrated in situated action. Both make it possible to ground an attention to the shared that goes beyond the transcript, although they do not show that adding images to an agent improves its quality. [18, 19]

We propose to record separately that an object was produced, was shown, was mentioned by the user and received a relevant confirmation. These are distinct states. A click does not prove understanding of the whole content either. This separation prevents the system's memory from automatically turning a rendered screen into shared knowledge.

### 3.5. Voice detection and prediction of turn yielding

Voice Activity Projection, by Ekstedt and Skantze, studies the prediction of future voice activity and turn-taking events. Its interest for this project is to separate the detection of an acoustic pause from the estimation of what may happen next. That prediction is not equivalent to reading an emotional intention. [20]

In the prototype, a pause should keep its uncertainty until there are sufficient signals. The system can combine acoustic information, syntactic completion, act type, task context and expressed preferences. The design should offer a simple alternative when the prediction fails: an effective interruption, a control to hold the turn or an explicit way of saying one has finished.

## 4. Operational definitions and decision models

The following equations have different statuses: equations (1) and (5) are operational definitions; equation (2) is a temporal idealisation of implementation; equation (3) derives from a simplified loss matrix; and equation (4) is a generic template for decision under uncertainty. None has been fitted or validated with our own experimental data. Their contribution is to make explicit the decisions and assumptions that an implementation will have to test.

### 4.1. Audible gap between turns

$$g_i = t^{\text{onset}}_{\text{agent},i} - t^{\text{end}}_{\text{user},i}. \tag{1}$$

In equation (1), times are expressed in milliseconds on a common clock. The agent's onset corresponds to audio played on the device; the user's end is determined on the acoustic record, not on the detector's later decision. A positive value is a gap without speech; a negative one indicates overlap. Brief acknowledgements, cooperative overlaps and invasive interruptions are labelled separately. Without that classification, an average mixes different behaviours.

The protocol will have to set which end of utterance corresponds to each agent onset, including restarts and brief acknowledgements. If no response appears before the end of the observation, absence of transition or censoring will be recorded according to the design; not a latency of zero. A capture failure is missing data. The distribution of observed gaps must be accompanied by the frequency of unobserved responses.

Record resolution, clock synchronisation, acoustic onset criterion and disagreement between annotators will also be documented. The variability of gaps between people is not the uncertainty with which a gap is measured. The metrological distinction between repeatability and reproducibility requires stating which conditions are kept and which change; it is adopted here to specify technical tests, without equating subjective experience with a physical magnitude. [21]

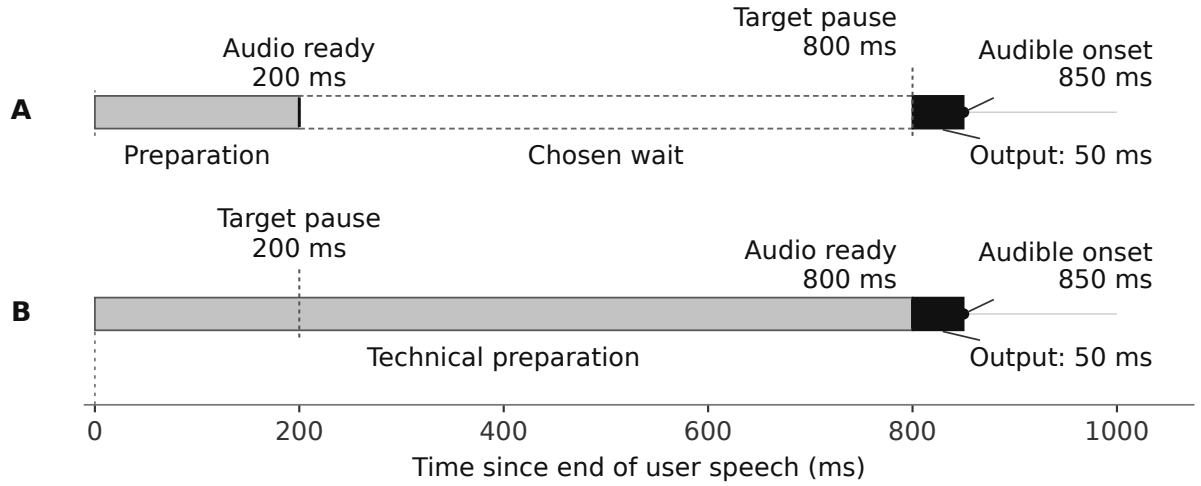



**Figure 1.** Two illustrative situations produce the same audible onset at 850 ms. In A, the target pause determines when to release the audio; in B, its technical availability does. In both cases the maximum of equation (2) applies, plus 50 ms of output. Hypothetical values: they do not come from a benchmark and are not product targets.

### 4.2. Technical availability and time of intervention

$$t_i^{\text{onset}} = \max\left(t_i^{\text{ready}},\ t_{\text{user},i}^{\text{end}} + \delta_i\right) + d_i^{\text{out}}. \tag{2}$$

Equation (2) describes the onset of a non-cancelled response turn, excluding brief acknowledgements. $t_i^{\text{ready}}$ is the instant at which the first playable fragment is available at a specified release point; $\delta_i \geq 0$ is the target pause and $d_i^{\text{out}} \geq 0$ the subsequent delay until it is audible. The timestamps $t_i^{\text{onset}}$, $t_i^{\text{ready}}$ and $t_{\text{user},i}^{\text{end}}$ refer to the same clock; $\delta_i$ and $d_i^{\text{out}}$ are durations. Any delay after the release point belongs to $d_i^{\text{out}}$, to avoid counting network or buffer twice.

Computation and deliberate waiting can overlap. That is why technical latency and pause must not be added blindly. If the audio takes 800 ms to be ready and the policy intended to wait 400 ms, the response does not show a decision to wait 800 ms: the infrastructure imposed at least part of the gap. The values are illustrative, not recommended targets. Figure 1 illustrates the distinction.

In a real implementation, $\delta_i$ is computed with estimates available online and may change while listening. The definitive acoustic end is known afterwards and serves to evaluate behaviour. If the user takes the turn back, the plan must be cancellable. The formula does not replace the listening policy.

### 4.3. Costs of interruption and delay

$$p_t > \frac{C_{\text{int}}}{C_{\text{int}} + C_{\text{del}}(\Delta t)}. \tag{3}$$

Let $F_t = 1$ be the operational label of turn yielding and $p_t = P(F_t = 1 \mid h_t)$ its estimated probability from the history. Two actions are considered: speak, or wait a fixed interval $\Delta t > 0$. Under the simplification of zero loss for speaking when there is yielding and waiting when there is not, the expected losses are:

$$L_{\text{speak}} = (1 - p_t)C_{\text{int}},$$
$$L_{\text{wait}} = p_t C_{\text{del}}(\Delta t).$$

Equation (3) follows from requiring $L_{\text{speak}} < L_{\text{wait}}$, with non-negative costs and a strictly positive sum. Both costs use the same unit of loss. If delay is expressed as a rate, $C_{\text{del}}(\Delta t) = \lambda \Delta t$, with $\lambda$ in loss per unit of time. In a tie we propose to keep listening. The rule is one-step and omits other costs and the value of future information that waiting could provide.

Costs will have to be estimated and revised by context; this article has neither fitted costs nor a calibrated probability. An implementation will have to define the reference label, document its uncertainty and evaluate calibration outside the fitting sample. During a hesitant story and before an urgent request, different errors may matter; that difference is a decision hypothesis, not an already known universal weight.

### 4.4. Selection of permitted actions

$$a_t^* \in \arg\max_{a \in \mathcal{A}_t} \mathbb{E}_{Z_t \sim b_t}[u(a, Z_t)]\,. \tag{4}$$

Here $b_t(z) = P(Z_t = z \mid h_t)$ is a distribution over interaction states relevant to the task, and $\mathcal{A}_t = \mathcal{A}_{\text{permitted}}(h_t)$ incorporates permissions and restrictions. $Z_t$ does not denote a directly observed psychological state. The action can be to speak, wait, acknowledge, clarify, offer or retrieve an object, propose saving it, repair or close. Membership of the argmax allows ties.

This expression is a decision template, not an estimated model. State, belief update, utility $u$, horizon and tie-breaking rule must be specified. Any combination of outcomes, effort and intrusion requires a common scale and declared weights; adding heterogeneous measures without that specification would lack operational meaning. Its validity will have to be examined through the evaluation of a concrete implementation.

Permissions restrict the set of actions before optimising it. A supposed engagement improvement cannot compensate for the lack of authorisation to keep a memory or share a file. The model must not maximise emotional exposure either. It can choose to end when the need has been met.

### 4.5. Measures of coordination and quality

$$R_{\text{premature}} = \frac{N_{\text{premature}}}{N_{\text{turns}}}. \tag{5}$$

In equation (5), $N_{\text{premature}}$ counts premature onsets and $N_{\text{turns}}$ agent onsets that could be coded; the numerator is a subset of the denominator. Each event receives a label according to a manual with signals of continuation, turn claim and, where appropriate, the participant's rating. Acoustic overlap alone does not determine the intention to speak.

Total, codable and missing onsets will be reported separately. With zero onsets the rate does not apply; if there were onsets but none could be coded, it cannot be estimated for lack of data. A pooled proportion weighted by turns will be distinguished from a per-person average. For stopping time, the onset of the interruption attempt and the criterion of audible cessation will be fixed; without an attempt, that measure does not apply. Median, percentiles and their uncertainties will be reported with their denominators and estimation method.

Quality is represented initially as a profile, not a total score: agreed outcome, temporal fit, understanding, continuity, control, effort and memory accuracy. PARADISE offers a precedent for relating conversational outcomes and costs to satisfaction through empirical estimation. It does not provide universal coefficients ready to transfer to this project. [22]

## 5. Technological capabilities and limits of comparison

### 5.1. Latency definitions and figures

Eleven v4 and Eleven v4 Turbo were announced on 28 September 2026. The catalogue describes a median inference of approximately 100 ms for Turbo, excluding application and network. This is speech synthesis, not the full interval between the user's last word and an agent's audible response. [23, 24]

A chain may include capture, recognition, end-of-turn decision, generation, tools, synthesis, network and playback. Some stages overlap through streaming. Observed latency must be instrumented on the whole system; adding isolated catalogue figures may count some times twice and omit others. ElevenLabs documentation distinguishes inference from time to first audio. [25]

The 150 ms TTS test used identical scripts, default settings and WebSocket streaming for v4 Turbo; it measured and discounted the network. The catalogue and announcement consulted do not provide sample size, dispersion or confidence intervals specific to the 100 and 150 ms latencies. They are kept as vendor statements under the described conditions. [23, 24]

The four Eleven Realtime and GPT-Realtime-2.1 figures in table 2 come directly from Artificial Analysis. Their TTFA is the mean time to the first output audio token on Big Bench Audio questions. It is an event prior to audible playback and must keep its definition. High and Minimal identify the evaluator's configurations; Minimal does not designate the Mini model variant. [26, 27]

The pages consulted do not provide effective $n$, repetitions, exclusions or confidence intervals per configuration for TTFA, and do not sufficiently formalise the timer's start event. The nominal size of the question set is not adopted as the effective number of measurements. The comparison is descriptive: the sources do not allow equal network, region, load, transport and all settings to be established, nor statistical significance or general superiority to be inferred. [27]

### 5.2. Modular and integrated architectures

ElevenLabs offers a Custom mode to choose components and a Realtime mode that manages recognition, model and synthesis. The consulted product announcement of Eleven Realtime v1 Mini and Max is dated 20 October 2026. The documentation describes a gradual rollout by account. Specific availability should be verified before designing an experiment that depends on that function. [28, 29]

OpenAI documents GPT-Realtime-2.1 as a speech-to-speech model with configurable reasoning, tool use and improvements in silence, noise and interruptions. Increasing reasoning effort can increase latency. Its model card accepts image input; that is not the same as generating images or deciding how they are shown in an interface. [30]

The choice of architecture depends on the control we need to inspect. A modular chain makes it easier to observe stages and swap components; an integrated solution can coordinate aspects of audio within a single system. This does not establish that one produces better atmosphere by definition. The prototype must compare concrete implementations and record their configurations.

### 5.3. Configuration of conversational behaviour

ElevenAgents allows interruptions and a readiness to respond to be configured through Eager, Normal and Patient. Expressive mode documents delivery control with instructions and tags, whose effect must be checked by voice and language. These controls can serve conversational direction, but a synthetic laugh or sigh does not show empathy. [31, 32]

OpenAI distinguishes server VAD, based on activity and silence, from semantic VAD, which estimates whether the user has finished from what was said. Configuration allows eagerness and response and interruption behaviour to be adjusted. The availability of these parameters does not remove the need to test real stories, hesitations and accents. [33]

*Design proposal.* Choose the manner of speaking after establishing what function the turn serves. A clarification may require a sober, brief voice; a story may

**Table 2.** Published figures by family of measure. TTS and TTFA correspond to different events and statistics.

| System | Published figure | What it represents | What it does not show |
|---|---|---|---|
| Eleven v4 Turbo | ≈100 ms median | Synthesis inference without application or network | The user's full wait |
| Eleven v4 Turbo | ≈150 ms median | Request to audible speech in a TTS test with network discounted | Interval from the end of the user's speech |
| Eleven Realtime v1 Max | 0.57 s mean | AA TTFA: first audio token, Big Bench Audio | Audible onset on the device |
| Eleven Realtime v1 Mini | 0.62 s mean | Same metric and evaluator | A statistically significant difference from Max |
| OpenAI GPT-Realtime-2.1 High | 1.21 s mean | Same metric and evaluator; High configuration | Universal performance of the model family |
| OpenAI GPT-Realtime-2.1 Minimal | 0.97 s mean | Same metric and evaluator; Minimal configuration | Equivalence with the Mini model variant |

*Note.* Sources: [23, 24, 26, 27]. AA: Artificial Analysis. The first two rows are TTS medians; the other four, the evaluator's TTFA means. They are not Divina Machina measurements. Consulted: 29 October 2026.

need prosodic variation. Do not use expressiveness as filler to disguise a poor answer or a tool still pending.

### 5.4. Consistency between generation, playback and state

OpenAI documentation distinguishes generated audio from played audio when there is an interruption. WebRTC and SIP allow the unplayed segment to be truncated automatically; over WebSocket the application must stop playback and communicate the corresponding point through `conversation.item.truncate`. [34]

Our design implication is broader: the shared history cannot contain as heard an explanation that was cut off. The same applies to an image that was never shown or a note whose saving failed. The system needs states of preparation, presentation and confirmation so as not to build continuity on events that did not happen.

Realtime also documents tools executed by the application. Continuity with asynchronous operations allows the conversation to continue while a result arrives, but does not authorise asserting it before receiving it. The product plan must distinguish pending, available, presented, accepted and cancelled. [35, 36]

## 6. Proposed architecture for situated conversation

### 6.1. Spaces of attention

Weiser and Brown proposed designing technologies capable of moving between the centre and the periphery of attention. Ishii and Ullmer explored objects, surfaces and ambient media as ways of relating digital information to situated activity. They are design antecedents that allow us to think of a conversational interface wider than a message history. They are not proofs of effectiveness for a current agent. [37, 38]

We propose three complementary spaces. The foreground contains what is being said or examined. A side space keeps objects available without demanding an immediate response. A working background prepares results, keeps temporal context and records pending operations. The person must be able to understand what is kept and bring to the foreground what they need.

The visual background must not become opaque storage. That a note does not interrupt the conversation does not mean it should be saved permanently and silently. The interface can show a discreet signal of preparation, allow a proposal to be reviewed at closing and keep only what the user has agreed to.

### 6.2. Visual objects and shared activity

The presentation policy can be formulated as a sequence. First identify what the object would serve for: comparing two forms, pointing to a detail, situating a memory or reviewing a decision. Then check whether it is available, whether its provenance is appropriate and whether the person wants to see it. Finally choose an entry point that does not compete with a turn still open.

The agent can say "I have a reference that may help compare that curve; I'll leave it to the side whenever you want". If the person changes topic, the object can expire without having been shown. Showing less can be better than interrupting every verbal association with visual content; this is a hypothesis to be tested, not a law.

Generated images require an explicit distinction between representation and evidence. If a scene is imagined, it must be presented as a proposal. If a personal photograph is retrieved, the agreed scope of use must be respected. No visual similarity authorises presenting an invented image as part of someone's real memory.

### 6.3. Presentation and circulation of objects

Sharing can mean placing an object within the session, creating a piece the person can download or sending it outside that session. We propose different actions and permissions for each case. An available file is not automatically sent to another person. The conversation can produce a draft, leave it pending and let the user decide its recipient and timing.

The object's state should distinguish preparation, availability, presentation, joint reference and persistence. If the user withdraws it, the application cancels future scheduled mentions. If they save it, it keeps the provenance and the version they actually saw. A link

that changes later is not equivalent to a stable copy of the shared object.

### 6.4. Memory, persistence and review

We propose three levels. Turn context keeps what is needed for the immediate intervention. Session memory sustains open topics and temporary objects. Cross-session memory keeps only authorised elements with an understandable purpose. These levels can implement different expiries and must not be confused with providers' technical retention.

Each persistent memory should have content, origin, date, scope, confirmation state and review mechanism. It must also distinguish what the person said from an interpretation by the system. Faced with a contradiction, the agent asks or updates; it does not force the new conversation to fit the earlier profile.

The policy can accept a stable authorisation, such as saving notes for a project, without interrupting to confirm every sentence. It must make that scope visible, offer exceptions and allow it to be reverted. For personal memories outside that agreement, an explicit proposal to save avoids turning listening into an automatic archive.

### 6.5. Editing conversational situations

The proposed unit of work is a situation: what the person wants to do, what activity they share, what information is missing, what restrictions exist and how a close could be recognised. Each situation admits several continuations. The editor defines resources and conditions; the actual words remain open to the interaction.

A team could describe a creative exploration situation with an available question, two optional references, a place to leave ideas and a clear way to finish. It could also describe an operational task with required data, checks and outcome. The system does not need to apply the same expressive intensity to both.

### 6.6. Temporal instrumentation of the experience

A viewer should align the user's audio, the agent's audio, turn decisions, operations, objects and memory events. The designer can listen to a gap and check what produced it: voice still being prepared, chosen wait, slow tool or wrong yielding. Interpretive annotations must be kept separate from the technical record.

The viewer needs to show discrepancies between what was generated and what was presented. A generated segment that was never played may remain in an authorised diagnostic log, but it is not recorded as presented. The segment actually played keeps its place in the history; its playback does not prove understanding either. A late result can be discarded because the topic changed. This makes it possible to study continuity without attributing to the user experiences that did not happen.

### 6.7. Coordination of turns and operations

We propose a coordinator above the voice adapters. It keeps the current purpose, the estimate of turn yielding, open topics, available objects, pending operations and permissions. It decides when to release a prepared response and when to postpone an image or a proposal to save.

The immediate temporal path must be able to work without waiting for slow work. If the person interrupts, the audio stops and the pending presentation is cancelled or marked obsolete. This does not revert an external operation already executed, whose state must be checked separately. If a tool is still working, its result is tagged with the version of the context that originated it. A result of an earlier intention must not reappear as if it were still current.

External operations require an additional policy. Preparing an alternative is reversible; sending a message or making a booking has effects. Anticipation can prepare data, but an action with effects must go through the permissions and confirmations that correspond to the product. Speculative generation does not constitute authorisation.

The separation between coordinator, voice adapters, object presentation and memory makes it possible to change provider without redefining all the interaction rules. That portability will have limits: each adapter must declare which events and controls it actually exposes, and the system must not invent missing data to complete its record.

### 6.8. Minimal event contract

Use monotonic clocks for local durations and document synchronisation between client and server. Store session, turn and model version identifiers. If the provider does not expose a probability, the field remains absent or is identified as an own estimate; a reconstructed figure is never attributed to the provider. Not all audio needs to be kept for all functions: the research protocol and the product must define separately which data they need and for how long.

### 6.9. Library of evaluation situations

The evaluation library should include a brief request, a person searching for a word, a story with an internal pause, a correction, an interruption during an explanation, a slow tool, a change of topic and a request to forget. Add conditions of noise, accents, speech rates and use without a screen. Do not automatically interpret a disfluency or an accent as difficulty in understanding.

Each situation needs an observable criterion. For example: allow a sentence to be completed without invasion; stop a response when interrupted; not claim that a note was saved if there was an error; remember only the confirmed version. The criteria can be partly automated, but ratings of atmosphere require research with people.

Table 3 summarises the events needed to instrument this architecture.

**Table 3.** Proposed minimal contract for the event log.

| Event | Minimal datum | What it is for |
|---|---|---|
| User's acoustic end | Time and annotation confidence | Measure gaps without depending on the detector |
| Estimated turn yielding | Time, configuration and probability if any | Explain the decision and identify own estimates |
| Audio ready and played | Two distinct timestamps and audible segment | Separate technical capacity from experience |
| Interruption and audio stop | User onset and end of playback | Measure effective return of the turn |
| Object ready and presented | Identifier, version and two times | Study deferred presentation |
| Memory proposed and saved | Scope, authorisation and outcome | Distinguish intention from persistence |

*Note.* Fields not exposed by a provider must remain absent or be identified as own estimates.

## 7. Illustrative application

The following example is original and fictitious. It illustrates design decisions and does not reproduce dialogue from the films or results of a study. The person is exploring an idea for an installation and has agreed to a slow-paced conversation mode. There is a side space for temporary notes; they have not yet authorised saving content across sessions.

The person says: "I want it to feel like going back to a place, but I don't know yet...". The sentence and its prosody suggest a continuation. The agent keeps listening even if there is a gap without voice. It does not present images or complete the sentence on the person's behalf.

The person continues: "...as if the place remembered you". The system prepares a short response and releases it when it estimates yielding. "What could it recognise in you without you having to explain everything?". The question offers an angle of exploration and allows an answer different from the expected one.

While the person develops the idea, the system can prepare two authorised visual references. It keeps them as available, without showing them yet. When the user asks for examples, it offers one image at a time. If the user points to a detail, the reference becomes part of the conversation. If they ignore it, the system does not presume understanding or insist because it produced it.

Later, the agent proposes a temporary note: "There's an idea that seems useful: recognising without asking for an explanation. I can write it down here". If the user accepts, it appears in the session. At closing, it can offer to save it for the project. Confirmation of saving happens after the operation has completed, with the option to edit or delete.

The scene must also work if the person says "No, that's not it". The agent withdraws its interpretation and asks to clarify only what is necessary. It does not defend its metaphor or try at all costs to recover a poetic atmosphere. Continuity consists of following what the person is building.

In an operational task the same coordinator would have another policy. Faced with "What time does the session start?", it would offer the verified datum directly. The quality of the design lies in recognising that difference of purpose, without turning every conversation into a contemplative experience.

## 8. Proposed research design

### 8.1. Building the film corpus

The next phase should identify the editions of the three films and annotate complete sequences. We propose starting with the record booth and café in Sunrise; café, car and transition to the apartment in Sunset; car, meal and hotel in Midnight. It is an initial analytical selection, not a representative sample of the whole trilogy.

The observable layers include speech per participant, gaps, overlaps, visible gestures, shot changes, movements, sound and objects. Another layer collects interpretations: function of silence, change of topic, opening, exposure and repair. The protocol foresees two coders working with a common manual, comparing agreements and keeping disagreements. Times would be annotated on the identified edition, considering cuts and possible differences in playback speed.

The result will be a catalogue of resources and questions. The films' gaps will not be used as a normative distribution of human speech, because script, acting, camera and editing contribute to the result.

### 8.2. Perception and interaction studies

A first study can use original scenes recorded for the project. Words, voice, prosody and objects are kept, modifying only certain gaps. Observers rate appropriateness, continuity and discomfort. Assignment of versions is random. This investigates the perception of a scene without attributing to listeners an interaction experience they did not have.

A second study requires participants who actually converse with an agent. Keeping model, voice, instructions and tools constant, additional delays are introduced at random after a response is available. Both the assigned condition and the effective delay must be recorded. The hypothesis asks whether the effect of time depends on the task, not whether a number of milliseconds always wins.

### 8.3. Comparison of policies and modalities

The next comparison sets a fixed turn-yielding policy against a contextual one. To separate adaptation from

**Table 4.** Research hypotheses, comparisons and conditions that would weaken them.

| Proposed hypothesis | Comparison | Evidence that would weaken it |
|---|---|---|
| A contextual policy reduces invasive entries | Fixed and contextual policy with comparable times | A precise estimate excluding the minimum relevant improvement, or evidence of deterioration |
| A relevant object can ease understanding | Agreed object, object at closing and audio only | Additional distraction and no relevant improvement estimated precisely |
| Confirmed memory can sustain continuity | Several sessions with and without persistent memories | Errors, irrelevance or lower perceived control |
| Speed works differently depending on context | Practical task and exploration with assigned delays | An interaction estimated precisely incompatible with the proposed minimum effect |

*Note.* Proposed hypotheses; the article presents no experimental results of its own.

speed, a condition with a matched global temporal distribution should be added. If the contextual policy is simply faster in every case, we will not be able to attribute the difference to it understanding the moment better.

Visual objects are studied afterwards, with a stable temporal policy: audio without object; the same relevant object presented at an agreed point; and that object available at closing. Memory is studied over repeated sessions: session context against confirmed cross-session memories, with editing and deletion in both cases where appropriate. Changing voice, latency, personality, images and memory at the same time would prevent identifying what produced an effect.

Comparison between providers is carried out as an evaluation of complete systems, with documented configurations. Changing provider can change voice, recognition and content, so it is not a clean experiment on the effect of latency. These two research aims must be kept separate.

### 8.4. Hypotheses and refutation conditions

Research must allow for part of the proposal not working. An evaluation whose only acceptable result is "more atmosphere" would not allow design to be distinguished from confirmation of an intuition.

An imprecise estimate or a non-significant result does not show the absence of a relevant effect. Table 4 must be interpreted against a minimum improvement specified before the analysis. If equivalence is to be concluded, its margins and the corresponding test will be defined beforehand; the data may also prove inconclusive. [39]

### 8.5. Evaluation variables and analysis

Each study will have to fix a main contrast, its population, task, horizon and a primary variable before collecting confirmatory data. For the turn-taking hypothesis we propose estimating the difference between policies in the probability of premature onset, expressed in percentage points. Purpose success, control, effort, understanding and continuity will be included according to an explicit plan of secondary measures or multiplicity. Ratings of atmosphere will remain exploratory until their reliability and structure have been studied.

Gaps, stopping after interruption and tool failures will be recorded, with device, network, language, noise and configuration. Denominators, losses, exclusions and unobserved responses will be reported; slow turns will not be excluded for making performance look worse. Temporal percentiles will require enough events and a declared method of uncertainty. Turns from one person or conversation will not be treated as independent participants.

The pilot will make it possible to review the instrumentation and estimate variability. The confirmatory size will be justified through a minimum relevant difference and a power or precision analysis that respects repeated measures. The plan will specify unit of randomisation, counterbalancing, order effects, stopping rule and handling of dependence between people, conversations and stimuli. The statistical family must correspond to the variable: for example, binary events or ordinal observations. Effects and uncertainty intervals will be presented, with predefined sensitivity analyses for missing data and coding decisions.

Session duration, number of turns or amount of personal information revealed are not proposed as primary indicators of success. They can increase through confusion or unnecessary questions. A brief and satisfying close can be an excellent result.

Table 4 makes explicit the comparisons that allow the hypotheses to be revised or rejected.

### 8.6. Initial development programme

We propose a sequence of three deliverables. The first is an atlas of situations, with corpus annotations and a vocabulary of actions. The second is an instrumented prototype that allows temporal policies to be changed, optional objects to be shown and session memory to be managed. The third is a pilot with people that checks whether those decisions produce differences and where they fail.

To start the prototype, two contexts suffice: a practical task with a verifiable answer and a brief creative exploration. A voice engine, an action coordinator and an event log are enough to test the main hypotheses. Cross-session memory can be added after verifying turn control and clarity of persistence.

An initial team could combine experience direction, voice and interaction engineering, and research support

in conversation analysis or HCI. The project's expected contribution would lie in the corpus to be built, the coordination policy, the editor, the instrumentation and the evidence its evaluation produces. The choice of provider is one part of that construction.

The criterion for extending development should be an observable improvement over a simple baseline: less invasion, better understanding or useful continuity without reducing control. If a visual module distracts or a memory adds no value, it is withdrawn or revised. The proposal must be able to show usefulness without depending on a specially prepared demonstration.

### 8.7. Preregistration and reproducible materials

Before the confirmatory study we propose registering hypotheses, main variable, minimum relevant difference, sample, assignment, exclusions, stopping and analysis plan. Later changes will be dated and planned analyses will be distinguished from exploratory ones. This procedure favours the traceability of decisions; it does not prevent exploring unexpected results. There is not yet a preregistration of the experiments proposed in this article. [40]

The materials will have to fix agent instructions, voice, model identifier and version, turn parameters, transport, region and tool configuration. The annotation manual will define onset, end, brief response, interruption and indeterminate case, with examples and a resolution procedure. Human evaluation will document who knows each condition; it will be acknowledged when the modality makes concealing it impossible. Data, code and materials will be shared to the extent compatible with participation agreements, with documented versions and changes.

## 9. Discussion and limitations

This document does not yet quantify the scenes, does not run provider tests and does not present a validated instrument of atmosphere. The review prioritises primary sources and directly relevant research, but does not exhaust conversation analysis, film studies or human-computer interaction. Commercial figures come from changing documentation and require new checks at implementation time.

*Current availability.* The LaTeX project includes the references, the code and the illustrative figure, together with a structured register of the central figures. That register contains transcribed statistics and their known conditions; it is not an experimental dataset. Divina Machina has not produced, in this work, latency measurements, temporal annotations of the corpus or results with participants. Reproducing the layout is possible with the delivered files; independent reproduction of each benchmark would require records and procedures that their sources do not always publish.

The transfer from a fictional couple to a relationship with an agent has structural limits. There is no equivalent human reciprocity; an expressive voice can produce attributions the system cannot sustain. That is why we propose clarity about the agent's artificial nature, an effective ability to leave and an understandable memory policy. Aesthetics does not replace accuracy or keeping to what was agreed.

We must also study differences of accessibility and context. A pause can help one person and create uncertainty in another; an image does not contribute the same with and without a screen. Expressed preferences and available controls must be able to prevail over automatic estimates. Variation is not noise to be removed from design.

## 10. Conclusions

Our reading of the *Before* trilogy places conversation in shared time: what is said acquires meaning together with what has just happened and what can still happen. That reading guides a technical proposal that distinguishes availability, intervention, presentation and persistence. Its usefulness will have to be established through the comparisons described in this article.

We can make explicit the rules by which a system takes the floor, offers an object, acknowledges a correction or keeps a memory. We can also measure its errors and ask people how they affected the experience. The precision of the record makes it possible to review the design without attributing to measurements a meaning they do not show.

The proposed criterion is that intervention should contribute to the purpose of the person conversing and preserve their ability to change it. A response, an image or a note can leave something available for later. On other occasions, the system's work will be done when the person finds their own words and no longer needs another question.

**Signed.** Cris Olmedo and Kurt Sommer, Divina Machina Research Group. *How to cite*: Olmedo, C. and Sommer, K. (2026). The art of conversation. Working paper DM-2026-01, v 0.2. Divina Machina Research Group. Divina Machina by Qs Ventures.

## References

Digital sources were consulted on 29 October 2026. Film links verify credits and context; details of staging rely on interviews and are distinguished from our readings. Undated documentation pages are cited as current documentation consulted. Commercial statements do not constitute an independent evaluation by this project.

[1] R. Linklater. Before Sunrise. The Criterion Collection, credits page, 1995. URL https://www.criterion.com/films/28692-before-sunrise. Film. Directed by R. Linklater; written by R. Linklater and K. Krizan.

[2] R. Linklater. Before Sunset. The Criterion Collection, credits page, 2004. URL https://www.criterion.com/films/28693-before-sunset. Film. Directed by R. Linklater; written by R. Linklater, J. Delpy and E. Hawke; story by R. Linklater and K. Krizan.

[3] R. Linklater. Before Midnight. The Criterion Collection, credits page, 2013. URL https://www.criterion.com/films/28828-before-midnight. Film. Directed by R. Linklater; written by R. Linklater, J. Delpy and E. Hawke.

[4] R. K. Osenlund. Interview: Julie Delpy and Richard Linklater on Before Midnight. Slant Magazine, May 2013. URL https://www.slantmagazine.com/features/interview-julie-delpy-and-richard-linklater/. Dated 23 May 2013. Primary interview.

[5] S. Bland. Richard Linklater on Before Sunrise's enduring legacy. Huck, September 2020. URL https://www.huckmag.com/articl

e/richard-linklater-on-before-sunrises-enduring-legacy. Dated 28 September 2020. Primary interview.

[6] Phil Hoad. Julie Delpy and Ethan Hawke: how we made the Before Sunrise trilogy. The Guardian, November 2019. URL https://www.theguardian.com/film/2019/nov/04/julie-delpy-ethan-hawke-how-we-made-before-sunrise-trilogy-sunset-midnight. Dated 4 November 2019. Interviews. Updated 6 November 2019.

[7] N. Minow. Interview with Richard Linklater about Before Sunset. More about Movies and More, July 2004. URL https://nellminow.blogspot.com/2004/07/interview-with-richard-linklater-about.html. Dated 8 July 2004. Primary interview.

[8] Time Out. TONY Q and A: Before Midnight's Richard Linklater, Ethan Hawke and Julie Delpy. Time Out, May 2013. URL https://www.timeout.com/film/tony-q-a-before-midnights-richard-linklater-ethan-hawke-and-julie-delpy. Dated 21 May 2013. Primary interview.

[9] M. Koresky and J. Reichert. A conversation with Richard Linklater. Reverse Shot, July 2004. URL https://reverseshot.org/symposiums/entry/204/richard-linklater. Dated 2 July 2004. Primary interview.

[10] P. Sobczynski. Interview: André Gregory and Wallace Shawn. RogerEbert.com, June 2015. URL https://www.rogerebert.com/interviews/interview-andre-gregory-and-wallace-shawn. Dated 12 June 2015. Primary interview.

[11] Ł. Mańkowski. On the road with Hamaguchi Ryūsuke. BFI Sight and Sound, November 2021. URL https://www.bfi.org.uk/sight-and-sound/interviews/hamaguchi-ryusuke-drive-my-car. Dated 16 November 2021. Primary interview.

[12] C. Aguilar. Embodiments of time: Kogonada on After Yang. RogerEbert.com, March 2022. URL https://www.rogerebert.com/interviews/embodiments-of-time-kogonada-on-after-yang. Dated 3 March 2022. Primary interview.

[13] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974. URL https://www.conversationanalysis.org/schegloff-media-archive/simplest-systematics-for-turn-taking-language-1974/. Accessed through the International Society for Conversation Analysis archive.

[14] T. Stivers et al. Universals and cultural variation in turn-taking in conversation. *PNAS*, 106(26):10587–10592, 2009. doi: 10.1073/pnas.0903616106. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC2705608/.

[15] S. C. Levinson and F. Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6:731, 2015. doi: 10.3389/fpsyg.2015.00731. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2015.00731/full.

[16] E. M. Templeton, L. J. Chang, E. A. Reynolds, M. D. Cone LeBeaumont, and T. Wheatley. Fast response times signal social connection in conversation. *PNAS*, 119(4):e2116915119, 2022. doi: 10.1073/pnas.2116915119. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8794835/.

[17] E. M. Templeton, L. J. Chang, E. A. Reynolds, M. D. Cone LeBeaumont, and T. Wheatley. Long gaps between turns are awkward for strangers but not for friends. *Philosophical Transactions of the Royal Society B*, 378:20210471, 2023. doi: 10.1098/rstb.2021.0471. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9985966/.

[18] H. H. Clark and S. E. Brennan. Grounding in communication. In *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association, 1991. URL https://web.stanford.edu/~clark/1990s/Clark%2C%20H.H.%20_%20Brennan%2C%20S.E.%20_Grounding%20in%20communication_%201991.pdf. Author copy at Stanford.

[19] C. Goodwin. Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10):1489–1522, 2000. doi: 10.1016/S0378-2166(99)00096-X. URL https://www.sciencedirect.com/science/article/pii/S037821669900096X.

[20] E. Ekstedt and G. Skantze. Voice Activity Projection: Self-supervised learning of turn-taking events. In *Interspeech 2022*, pages 5190–5194, 2022. doi: 10.21437/Interspeech.2022-10955. URL https://www.isca-archive.org/interspeech_2022/ekstedt22_interspeech.html.

[21] Barry N. Taylor and Chris E. Kuyatt. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical Note 1297, National Institute of Standards and Technology, 1994. URL https://www.nist.gov/pml/nist-technical-note-1297/nist-tn-1297-appendix-d1-terminology. 1994 edition. Appendix D.1, Terminology, web version.

[22] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *ACL EACL*, pages 271–280, 1997. doi: 10.3115/976909.979652. URL https://aclanthology.org/P97-1035/.

[23] M. Staniszewski and P. Dabkowski. Introducing Eleven v4: our most emotive model. ElevenLabs, September 2026. URL https://elevenlabs.io/blog/eleven-v4. Dated 28 September 2026, updated 19 October 2026. Methodology note 3 documents the 150 ms figure.

[24] ElevenLabs. Models. ElevenLabs documentation, n.d.. URL https://elevenlabs.io/docs/overview/models. Models, identifiers and inference latencies.

[25] ElevenLabs. Understanding latency. ElevenLabs documentation, n.d.. URL https://elevenlabs.io/docs/eleven-api/concepts/latency. Inference, first audio and system stages.

[26] Artificial Analysis. Speech to speech models and providers analysis. Artificial Analysis, n.d.. URL https://artificialanalysis.ai/speech-to-speech. Live results table; explicit model configurations.

[27] Artificial Analysis. Speech to speech benchmarking methodology. Artificial Analysis, n.d.. URL https://artificialanalysis.ai/methodology/speech-to-speech-benchmarking. TTFA definition and Big Bench Audio description.

[28] M. Kim. Introducing Eleven Realtime: a simpler way to build natural voice agents. ElevenLabs, October 2026. URL https://elevenlabs.io/blog/realtime. Dated 20 October 2026. Commercial launch report; benchmarks attributed to Artificial Analysis.

[29] ElevenLabs. Realtime models. ElevenLabs documentation, n.d.. URL https://elevenlabs.io/docs/eleven-agents/customization/realtime-models. Custom, Realtime, settings and gradual availability.

[30] OpenAI. GPT-Realtime-2.1 Model. OpenAI documentation, n.d.. URL https://developers.openai.com/api/docs/models/gpt-realtime-2.1. gpt-realtime-2.1 model card, modalities and reasoning effort.

[31] ElevenLabs. Conversation flow. ElevenLabs documentation, n.d.. URL https://elevenlabs.io/docs/eleven-agents/customization/conversation-flow. Timing, interruptions and turn eagerness.

[32] ElevenLabs. Expressive mode. ElevenLabs documentation, n.d.. URL https://elevenlabs.io/docs/eleven-agents/customization/voice/expressive-mode. Expressive control and output variation.

[33] OpenAI. Voice activity detection. OpenAI documentation, n.d.. URL https://developers.openai.com/api/docs/guides/realtime-vad. Server VAD and semantic VAD.

[34] OpenAI. Realtime conversations. OpenAI documentation, n.d.. URL https://developers.openai.com/api/docs/guides/realtime-conversations. Conversation management, interruptions and audio truncation.

[35] OpenAI. Realtime with tools. OpenAI documentation, n.d.. URL https://developers.openai.com/api/docs/guides/realtime-mcp. Functions and tools in Realtime.

[36] OpenAI. Developer notes on the Realtime API. OpenAI, n.d.. URL https://developers.openai.com/blog/realtime-api. Reference used for asynchronous calls.

[37] M. Weiser and J. S. Brown. Designing calm technology. Xerox PARC, December 1995. URL https://calmtech.com/papers/designing-calm-technology. Dated 21 December 1995. Reproduction of the original text on Calmtech.com.

[38] H. Ishii and B. Ullmer. Tangible Bits: Towards seamless interfaces between people, bits and atoms. In *CHI 97*. ACM, 1997. URL https://chi1997.acm.org/proceedings/paper/hi.html. Digital proceedings.

[39] Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5502906/.

[40] Center for Open Science. Preregistration. Center for Open Science, n.d. URL https://www.cos.io/initiatives/prereg.