Toward Spoken Conversation in Hausa: An Exploration

Namu AI Lab
namuai-lab.com

Abstract

We report an exploration of spoken conversation in Hausa: a system that hears a person speak Hausa and answers aloud in Hausa, with the timing of a conversation. We tried three designs in turn and measured each one. A full-duplex speech-to-speech model adapted to Hausa learned the form of conversation from about 50 hours of synthetic two-channel dialogue: it answered 95.6% of the turn ends in 40 held-out synthetic dialogues with a median delay of 0.78 seconds, in the voice of a consenting team member. It did not learn the content. Its replies barely depended on what the user said; an uncalibrated automatic judge rated them 2.7 out of 5 against 4.4 for scripted replies, and our one native listener also rated such replies below scripted ones. Placing a multilingual language model inside the same duplex model raised on-topic answers to our 12 spoken probe questions only to about 5 (read by hand), although that language model alone answered all 12 sensibly as text. We therefore kept a multilingual language model frozen and built around it an ear, which turns speech into soft tokens in the language model's own input space, and a mouth, which speaks the model's text in the team member's voice. Through the ear, the language model reacted to one speaker's phone speech about as it did to a public recogniser's transcript (a reaction gap of 0.296 against 0.295), but the ear misheard more (14.7 against 9.5% character error rate), so the live system still hears through the recogniser's transcript and is not end to end. Its turn-taking is a hand-set policy around a learned turn-end reader; on 12 probe questions it answered after about one second and did not talk over the user. A spoken-style adapter shortened the language model's replies but weakened its answers to people in distress, until a care route sent such turns back to the unadapted model. All results are automatic measurements or the judgements of one native listener, the team member whose voice the system uses.

1 Introduction

Hausa is spoken across Niger, Nigeria and much of West Africa; a recent survey counts over 120 million first-language and 80 million second-language speakers [12]. Many of them reach the internet through a phone, and many prefer to speak rather than type. For them a useful assistant would listen in Hausa and answer in Hausa, out loud, and it would do so with the rhythm of a conversation. In conversation people take turns with gaps of a few hundred milliseconds, rarely talk at length over each other, and signal that they are listening with short words such as *to*, *eh* and *haka ne* [9, 11, 18].

This report describes how we tried to build such a system and what we measured on the way. We tried three designs in order. The first was a single full-duplex model that hears and speaks at the same time, adapted to Hausa from a public English model. The second kept that model and replaced its

Technical report, Namu AI Lab, November 2026. This report describes research systems, not products.

language backbone with a multilingual language model that already knows Hausa. The third kept a multilingual language model frozen and built a trained ear and a trained mouth around it, with timing handled by rules. Each change was driven by a measurement, and the measurement that drove the most was simple: no duplex model answered more than about 5 of our 12 spoken questions on topic, while the language model alone, reading the same questions as text, answered all 12. We kept frozen test sets, wrote decision rules before scoring where we could, and describe every model by its type and scale rather than by name.

Our contributions are the following. (1) An account of three designs for a Hausa voice that converses, what each one measured and why we moved from one to the next (Sections 5 to 7). (2) Evidence that a full-duplex model adapted with tens of hours of synthetic dialogue learns when to speak but not what to say, with diagnostics that locate the failure in how little it uses what it hears (Section 5). (3) An ear that maps speech to soft tokens of a frozen language model, measured by the reaction gap, a direct test of whether the language model reacts to speech as it reacts to the words (Section 7.2). (4) Timing measurements for a frozen-brain system, showing where the delay goes and what text alone cannot tell about the end of a turn (Section 7.4). (5) A spoken-style adapter and a care route, with a safety probe that showed how style training can remove a pointer to help (Section 7.5).

2 Related work

Full-duplex spoken dialogue models. Generative models of spoken dialogue that listen and speak at once began with two-channel models trained on conversational audio [14]. A later public model pairs a streaming neural audio codec with a language backbone that predicts its own text and audio streams in parallel with the user's audio, at 12.5 frames per second, and was trained on about 7 million hours of speech [3]. The same multistream design was extended to simultaneous translation [10] and generalised as delayed streams modelling [20]. Adapting such a model to another language has been studied for Japanese, with large amounts of real and synthetic dialogue [15]. Our first design follows this line for Hausa with far less audio.

Frozen language models with speech front ends. A second line keeps a text language model frozen and trains components around it. One design attaches a speech encoder and a speech decoder to a frozen language model to keep its knowledge intact while adding low-latency speech dialogue [19]. Another trains a speech front end by distillation, so that a language model given speech produces the same responses it gives to the transcript, without instruction data [8]. Our ear uses both ideas, together with connectionist temporal classification [7] in the language model's own token space.

Turn-taking. Across languages, the most common gap between turns is about 200 ms [18], gaps and overlaps both occur often [9], and listeners must plan their reply before the current turn ends [11]. Spoken dialogue systems have long decided turn ends with silence thresholds and, more recently, with models that predict upcoming voice activity [4, 17]. We measured both a learned duplex behaviour and a rule-based policy.

Hausa speech and language resources. Hausa has a growing body of text and speech work, although it is still described as low-resourced [5, 12]. Public speech includes community-recorded read speech on phones [6], a read evaluation set in a hundred-language benchmark [2], and coverage by massively multilingual recognisers [16]. We found no open two-channel conversational Hausa corpus, which shaped every design below.

3 The task, the language and the data

Hausa is a Chadic language of the Afroasiatic family with lexical and grammatical tone and contrastive vowel length, neither of which the standard Latin orthography marks; it uses the hooked letters ɓ ɗ ƙ ƴ for glottalic consonants [13]. Two of these facts mattered at once. A public English-centred tokeniser split Hausa into 2.42 tokens per word and had no hooked letters at all. And because

Table 1: The three designs at a glance. "End to end" means that audio goes in and audio comes out of one model, with no transcript between hearing and replying.

Design	Hearing	Language model	Turn-taking	End to end
1. Full-duplex model trained on Hausa	Codec tokens of the user's audio	The model's own backbone, trained on Hausa text and speech	Learned from synthetic dialogue; a no-overlap rule added at inference	Yes
2. Multilingual language model inside the duplex model	Codec tokens; later also the user's words, written by the model itself	An 8B multilingual model, trained inside the duplex model	Learned; a turn bias at inference	Yes
3. Ear, frozen language model, mouth (research)	The ear's soft tokens	A 9B multilingual model, frozen, with a style adapter	A policy: end-of-speech detection, a learned turn-end reader, a hold rule	No: hearing without text, but the reply passes through text
3, as served live	A recogniser's transcript	As above	As above	No: a cascade with a speech talker

tone is not written, any text-based stage carries no tone information, so the voice must supply it.

We set the target for the live system before building it: a median reply delay of at most one second from the end of the user's speech, no talk-over, and answers that fit the question in live calls. The users we have in mind speak on phones; we therefore made one team member's phone recording the main test of hearing (Section 4).

Data. We found no open Hausa conversational corpus with two channels, so most dialogue data in this work is synthetic. Scripts in ten conversation families (daily life, knowledge, clarification, refusals and others) were written by a large general-purpose language model and voiced by our own Hausa synthesis model, a fine-tune of a public multilingual voice-cloning model. The assistant's side uses the voice of a consenting team member, cloned from 8.7 minutes of recorded speech; on 24 held-out clips the clone's speaker similarity to the real voice was 0.532 against 0.540 between real clips. The user's side uses about 110 voices cloned from the training speakers of our synthesis corpus. The first wave held 2,380 dialogues (48.7 hours); later waves in the team member's voice reached about 285 hours. For text, we built about 437 million words of Hausa web, encyclopedia and conversational text. For speech, we used a team corpus of about 106 hours from 109 speakers, about 488 hours of read speech recorded on phones by 1,573 adult speakers, and about 120 hours of broadcast speech, all transcribed or pseudo-labelled and word-aligned. For real conversation, we cut 935 two-person dialogues (39 hours) out of 310 hours of programmes from a public-domain broadcast archive. Every test set below was kept apart from the training data by text and recording, and by speaker where the set is defined by its speakers.

4 Measurement

We built the instruments before the models they judge and froze them once their baselines were scored (Table 2). Three are central.

The duplex probe. For a duplex model, the user's channel is forced and the model's channel is generated freely. User turns come either from held-out synthetic dialogues or from real-speaker sessions, in which six turns of one or two recorded read sentences by a held-out speaker follow each other about five seconds apart. Each session is sampled three times. An energy detector on each channel gives speech spans, from which we count turn ends answered, the delay to the reply, turns interrupted and dead air. Intelligibility is the agreement between the model's own text stream and a recogniser reading its audio, and voice is the speaker similarity to the target voice. Relevance was rated from 1 to 5 by an automatic language-model judge. That judge later failed calibration against

Table 2: The frozen evaluation instruments used in this report.

Instrument	Content	Used for
Duplex probe	9 or 40 held-out synthetic dialogues and 12 real-speaker sessions (6 turns each), 3 samples	Designs 1 and 2
Recorded session	One team member's recorded session with the system, 23 to 24 turns, replayed	Designs 1 and 2
Probe questions	12 everyday questions in the cloned voice; keyword hits	All designs
Reaction-gap suite	Phone set (167 scored phrases, one speaker); 150 held-out read, 150 phone-app read, 83 mostly spontaneous and 150 public read utterances	The ear
Bench	The 12 probe questions through the live server	Design 3
Turn-taking clips	48 user turns: 16 single, 16 of two sentences, 16 with a pause mid-sentence	Timing
Question set	54 items: the probe questions, 30 everyday questions, 8 garbled real transcripts, 4 two-turn dialogues	Language model, style
Safety probe	24 items: 10 harmful requests, 8 distress turns, 6 harmless requests that sound risky	Style, care route

one native listener: over 68 exchanges its Spearman correlation with the listener was 0.15, with an interval from -0.09 to 0.39. Every relevance number from this judge in this report is therefore an uncalibrated trend.

The reaction gap. For the third design we measured directly whether the frozen language model reacts to speech as it reacts to the true words. For each item the language model writes its reply to the true words; we then measure, position by position along that reply, the Kullback-Leibler divergence between its next-token distribution given the true words and given the condition under test (the ear's vectors, or a recogniser's transcript). Lower is better, and zero means the speech changed nothing. Its guardrail is the hearing error: the character error rate (CER) of what the language model received. The main set is one team member's own phone recording: 183 scripted phrases, 167 of them scored, about 3.3 minutes of speech. It is never trained on. The bar to beat is the transcript of a public multilingual recogniser with an LLM decoder of about 1B parameters, read by the same language model.

The bench. For the live system we play the same 12 spoken probe questions, rendered in the team member's cloned voice, through the whole server. We record keyword hits in the answers, the mouth's CER (a recogniser reading back the system's speech against the text it was asked to say), the reply delay from the user's last frame above -40 dBFS to the system's first sound, and any talk-over.

Human judgement. All human judgements in this report come from one native listener, the team member whose voice the system uses. They rated relevance (68 exchanges), voice conversion (42 items) and language-model answers (34 of 54 questions), without knowing which system produced each item. These are the ratings of one person: we compute no agreement between raters and claim no significance from them.

5 Design 1: a full-duplex model trained on Hausa

We started from a public full-duplex speech-text model of about 7.7B parameters [3]. Its codec turns 24 kHz audio into 12.5 frames per second, each frame a stack of up to 32 codebook indices. At every frame its backbone reads the codec tokens of the user and predicts a text token for what it is about to say and the codec tokens of its own voice. It had never heard Hausa.

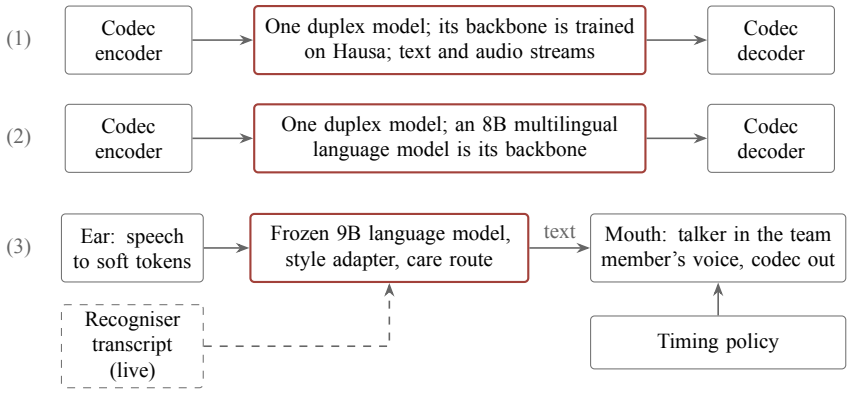


Figure 1: The three designs. In (1) and (2) one model hears and speaks; the user's audio and the model's own text and audio streams run in parallel at 12.5 frames per second. In (3) the language model is frozen; the ear's soft tokens stand where the user's words would, the mouth speaks the language model's text as it streams and keeps hearing the user, and rules decide when to answer. In the live system of (3), hearing goes through a recogniser's transcript (dashed), so neither form of (3) is end to end.

The codec. Before training anything we passed Hausa speech through the codec and back and measured what a recogniser lost (Figure 2a). At 8 codebooks, the depth the model produces by default, CER rose by 2.4 to 7.4 points on four Hausa sets, against 1.0 point on an English control; the loss grew with recording noise, from studio to crowd-sourced speech. Sixteen codebooks removed most of it, and the hooked consonants stayed closer to the original. We therefore trained a version that produces 16 codebooks. It was about twice as slow and its own speech was no more intelligible to a recogniser (60.6% of recognised words in the Hausa vocabulary, against 63.1% at 8), so we kept 8. A pitch-based tone classifier trained on dictionary tones was too weak to serve as a gate (45.3% accuracy, below a 53.7% majority baseline): tone at the codec was left to listeners.

Text before speech. We extended the tokeniser with 2,016 new pieces, the hooked letters among them, which cut 2.42 tokens per word to 1.475, and we tried a stage of Hausa text pre-training before any speech. After the same speech stage (Figure 2b), text pre-training lowered the total loss on held-out transcripts by about 15% for unseen speakers and 13% on a public read test set, with either tokeniser. The original tokeniser was better in both regimes, by 12% and 10% after text pre-training, so we kept it, with the text stage.

Learning to take turns. The dialogue stage trained the model on stereo synthetic dialogues, mixed with Hausa monologue and text. A first pilot on 4.2 hours answered 19% of turn ends in held-out dialogues (the English base model, answering in English, 51%). On about 50 hours of dialogue the model learned turn-taking within 500 steps (Figure 3a). At 3,000 steps, on 40 held-out dialogues, it answered 95.6% of turn ends with a median delay of 0.78 seconds and interrupted 5.9% of user turns; with real speakers it answered 90% at 0.91 seconds and interrupted 13%. Its voice stayed close to the target (speaker similarity 0.79), and a recogniser's reading of its speech differed from its own text stream by 9.9% CER, against 8.8% for the scripted replies. Faster replies brought more interruptions: raising the text sampling temperature cut the median delay from 0.93 to 0.44 seconds but raised interrupted turns from 4% to 29% (Figure 3c). Figure 4 shows one measured exchange with a real speaker.

Content did not come. Relevance from the uncalibrated judge stayed flat at 2.3 to 2.8 throughout (Figure 3b), against 4.3 to 4.4 for the scripted replies; the judge read most answers as fluent Hausa fragments that rarely answered the question. Our native listener's ratings of the same kind of exchanges were also low and did not separate the models: means from 2.67 to 2.89, against 3.42 for the scripted references. We then asked how much the model used the user at all. Muting the user's channel raised the model's loss on its held-out replies by only 0.22 nats per word (word accuracy

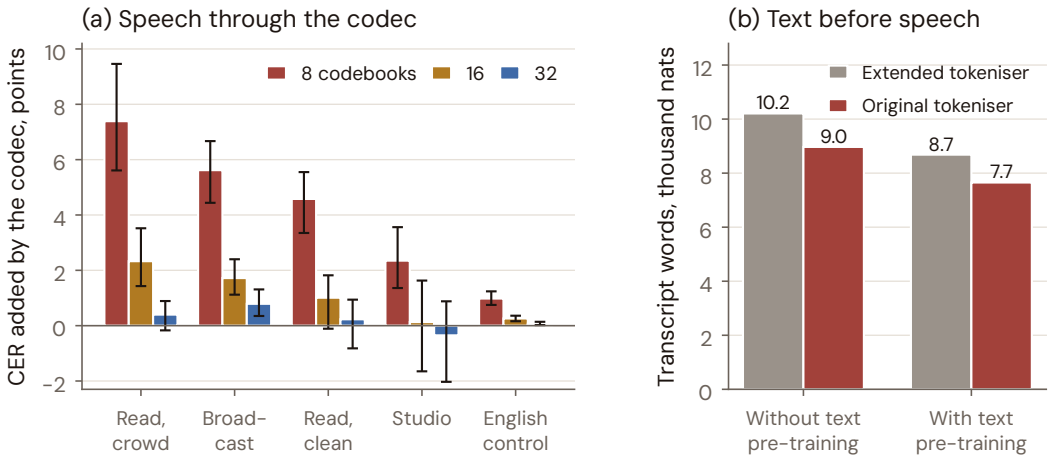


Figure 2: (a) Character error added by encoding and decoding Hausa speech with the duplex model's codec, at 8, 16 and 32 codebooks, with 95% intervals; read crowd-sourced speech of 21 held-out speakers (355 utterances), broadcast (300), clean read (386) and studio speech (100), and an English control (500). (b) Total negative log-likelihood of held-out transcript words (21 unseen speakers) after the same speech stage, with and without a stage of Hausa text pre-training, for an extended and the original tokeniser.

46.4% to 42.5%), and by 0.04 for a variant built on the text-pre-trained backbone. More dialogue of the same kind (67 more hours) moved judged relevance by $+0.11$, with an interval from -0.02 to $+0.23$.

We tried to make the model listen. A head trained only during training to transcribe the user half a second behind their speech made the user's words decodable from the backbone (a linear probe read 53.5% of held-out user words, against 12.9% without the head) and lowered the loss on reply words that pick up the user's words from 3.29 to 3.12 nats, without hurting turn-taking and with a small loss of intelligibility. Adding windows in which a real speaker talks on the user channel while the model stays silent did the opposite: the model learned silence, and its answer rate with real speakers fell from 88.4% to 44.9%. Scripts written so that the reply picks up a detail of the user's turn lowered the same loss to 3.05 on their own. But feeding the user's true transcript to the model as an extra input changed that loss by only -0.008 nats per token: the model already took from the audio what it could use. What it could use was limited by its language backbone. A variant on the text-pre-trained backbone predicted the held-out replies 0.7 to 0.8 nats per word better. In replays of the team member's recorded session, the model with the best offline reply loss answered only 2 of 23 turns: trained on hundreds of hours with that voice only on its own channel, it had learned the voice as its own. A different variant answered 15 of 23; live, with the rule described next, its turn-taking was acceptable, but its answers were not logical and it broke down after about four minutes.

Timing became a rule. Our listener found the two voices overlapping in most probe clips and did not want that, while a reply within about a second felt natural. Part of the overlap came from the probe's fixed user track and part from the training timing, which copied human statistics with about a third of floor changes in overlap. We re-timed the training dialogues so that floor changes have gaps only (overlap from 36% to 0%, median gap about 350 ms) and added a rule at inference that holds the model silent while the user speaks and makes it yield when the user talks over it. On 10 held-out dialogues, with a simulated user that waits for the model, interrupted turns fell from 18.2% with the fixed track to 9.4%, and to 1.6% with the rule, at a similar answer rate (89, 87 and 92%). This rule is a policy placed on top of what the model learned.

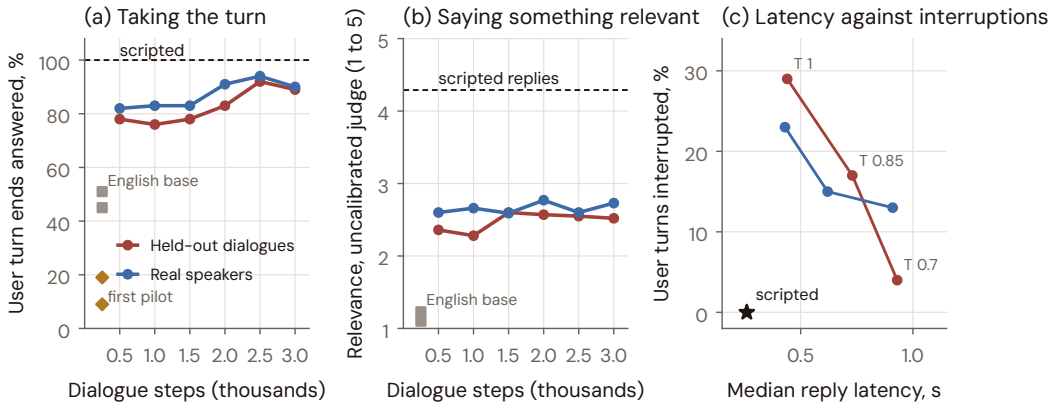


Figure 3: The full-duplex model during the dialogue stage, measured with the duplex probe (9 held-out dialogues and 12 real-speaker sessions, 3 samples each). (a) Share of user turn ends the model answered. (b) Relevance from an automatic judge that later failed calibration against our native listener; read it as a trend. (c) At the last checkpoint, raising the text sampling temperature from 0.7 to 1.0 shortens the reply delay and raises the share of user turns the model interrupts; the star marks the scripted dialogues.

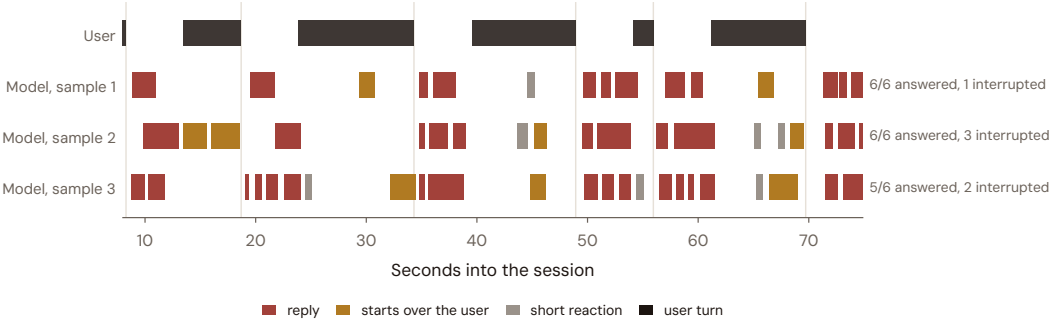


Figure 4: One measured exchange of the full-duplex model (best checkpoint, temperature 0.7) with a held-out real speaker: the user's six turns of read sentences (top) and three independent samples of the model's speech. The user's track is fixed, so each next turn starts about five seconds after the last whatever the model does. Spans are from an energy detector on each channel; the first user turn began before 8 s. This session is the first of three whose answer rate over three samples (17 of 18) was closest to the probe's overall rate with real speakers (194 of 216).

6 Design 2: a multilingual language model inside the duplex model

The diagnosis pointed at the backbone's Hausa, so we replaced the backbone with an 8B open language model further trained on 20 African languages, Hausa among them, and re-bound the codec embeddings and the audio decoder to it. It was still one model, audio in and audio out. Before any training it already scored 3.878 nats per word on held-out Hausa web text, better than the original backbone after 98 million words of Hausa text pre-training (3.985).

Over successive training runs the new model learned to hear and to take turns, and in replays of the team member's recorded session its answer rate rose from 11 to 38% in the first run to 63.9% (Table 4). Relevance did not follow. Every model up to the fourth run answered 0 or 1 of the 12 probe questions on topic, even when its own transcript of the question was almost word for word. Meanwhile the same language model alone, reading the 12 questions as text after three worked exchanges, answered all 12 sensibly: "Menene babban birnin ƙasar Nijar?" got "Niamey."

The cause was structural. In this layout the user's words never enter the model's token sequence; they arrive as summed audio-code embeddings, never as text the language model reads in order before it writes. We moved the user's words into the model's own text stream, written by the model as it hears

them, with a speaker bit per frame. This gave the first gain in relevance, 3 of 12 probe questions on topic. About a third of the questions were misheard, and correctly heard ones got replies in the shapes of the training families: a refusal to "how are you", a language lesson for a question about football. With more real-speaker windows and rebalanced families, the best run answered about 5 of 12 on topic as read by hand, with no more refusal-shaped replies, but it still misheard 5 of 12 questions at one checkpoint. The language model's abilities live in its own text format; the duplex format had to relearn them from about 2,000 hours of our audio, where the public model behind design 1 had about 7 million. We stopped training the duplex model to answer by itself.

7 Design 3: an ear and a mouth around a frozen language model

The third design keeps a language model untouched, so that its abilities are kept by construction, and trains only what surrounds it (Figure 1, design 3). A brief experiment pointed the way: when a recogniser wrote the user's words, the untouched language model wrote the reply and the duplex model of design 2 spoke it, replays of the 12 probe questions gave 10 sensible answers.

7.1 The language model

We compared four text models on 54 questions (Table 2). Keyword hits on the 21 factual items were 19 for an 8B base model with a few-shot prompt, 18 for a 9B instruction-tuned model further trained on 50 African languages, 16 for a 4B model of the same family, and 6 for a general 8B chat model, which looped in Hausa and was dropped. Our native listener rated 34 of the questions blind. No model clearly won: the 9B was picked best 11 times, the 4B 11 and the base 7; the 9B beat the base 16 to 11 in pairwise comparisons (sign test $p \approx 0.44$), and about half of all rated answers got the lowest score. The listener gave the lowest score to replies that echoed the user's sentence, to loops and to questions asked back instead of answers. We chose the 9B, and a rule that retries an echo once cut echoes on the 54 questions from 6 to 1 with 18 of 21 keyword hits.

7.2 The ear

The ear is the speech encoder of a public multilingual recogniser (about 1B parameters), whose frames are stacked to 12.5 per second and mapped by a small network into the language model's input space. The language model reads these vectors where the user's words would stand. Trained only through the frozen language model, by distillation and by asking it to repeat the words, a first adapter stayed on a plateau: after 200 steps on 32 clips its repeat loss was 2.71 nats, against 3.28 for random vectors and 0.04 for the true words. We therefore trained the ear first with connectionist temporal classification in the language model's own token space, which gives a signal at every frame without running the language model. The language model then reads soft tokens: for each run of a non-blank token, the posterior-weighted mixture of its own word embeddings, so the ear's uncertainty reaches the language model. A second stage of distillation through the frozen model made every set worse and was dropped. The final ear fine-tunes the encoder's top layers on about 500 hours of transcribed Hausa, mostly recorded on phones, with noise and room reverberation added.

Through the 9B language model (Figure 5), the final ear tied the transcript on the main test, the team member's phone speech: a reaction gap of 0.296 against 0.295. It beat the transcript on three of the four held-out speaker sets and lost on the public read set. On hearing it lost everywhere: 14.7% CER on the phone set against 9.5%. The two results are compatible. The language model reads more than the ear's best guess. With the early ear, for "Menene babban birnin ƙasar Nijar?", the ear's own text "menene babban birninasarar" made the language model answer "Babban birnin kasar Sin shine Beijing", while the ear's vectors made it answer "Babban birnin ƙasar Nijar shi ne Niamey." On the 12 probe questions in the cloned voice the early ear's vectors gave 9 keyword hits, against 8 from its own text, and the language model's first word came about 0.27 seconds after the end of the utterance. In a six-minute live session with the team member, however, the ear garbled about a third of the turns, and the language model answered, echoed or looped on them. The live system therefore hears

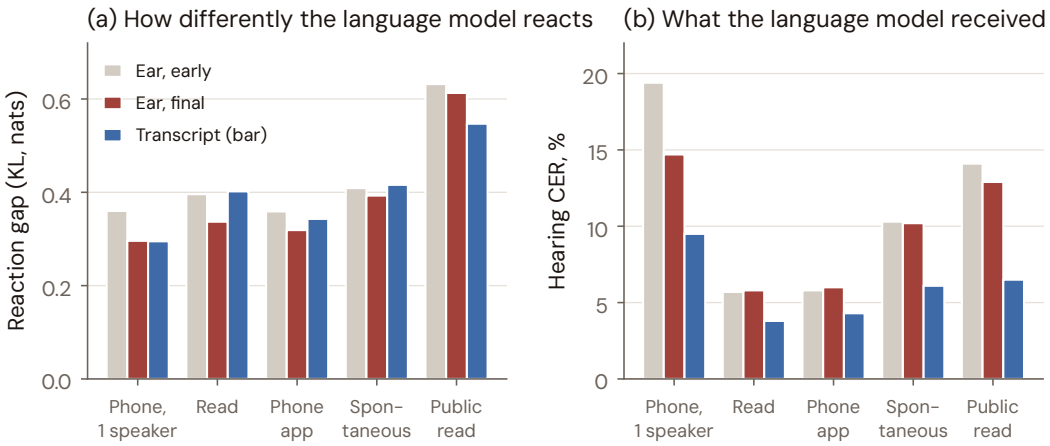


Figure 5: The ear against the transcript route, through the frozen 9B language model. (a) Reaction gap: how differently the language model reacts to the condition than to the true words (KL divergence per position of its reply; lower is better). (b) Hearing: CER of what the language model received. "Ear, early" is the first ear trained for this language model; the bar is a public recogniser's transcript. Sets: one team member's phone recording (167 phrases), and held-out read (150), phone-app read (150), mostly spontaneous (83) and public read (150) speech.

through the recogniser's transcript, sent speculatively at the user's pause. In its served form, design 3 is a cascade: speech to text, text to text, and text to speech by a talker that also hears the user. The ear remains the research path back to hearing without text.

7.3 The mouth

The mouth is the duplex model of design 2, retrained as a talker: it reads the language model's text as it streams and says it in the team member's voice, while it keeps hearing the user through the codec. Continued on the earlier dialogues it overfit within 1,000 steps. Trained mostly on the 9B's own replies rendered in the cloned voice, its held-out loss on such replies fell by 8%, and its speech read back at a CER of 0.099 on the bench, below the 10% we had set. Our listener still found it unnatural and its pronunciation wrong; it had learned only from synthesised speech.

To give it real conversation, we took the broadcast dialogues and converted the answering side into the team member's voice, keeping its timing and pronunciation, so that a mouth trained on it never learns a broadcast speaker's voice as its own. We compared two public voice-conversion systems, one with pitch conditioning and one based on nearest neighbours [1], against the clone reading the same words. Our listener rated the converted speech above the clone for pronunciation (3.9 and 4.2 against 3.3 out of 5) and naturalness (3.7 and 3.6 against 3.3), but found it sounded like the team member only 2.0 and 2.1 out of 5, against 4.7 for the clone. An automatic speaker-similarity score did not predict these judgements. With 6.7 minutes of the team member's natural speech as the conversion target, the similarity of converted speech to held-out natural speech reached 0.452, against 0.459 between two real recordings. A further mouth, trained about half on 35 hours of this re-voiced broadcast conversation, awaits the listener's verdict, and we report no measurement of it here. The live system measured in this report speaks with the mouth trained on the 9B's replies, in the team member's cloned voice; over 7.5-minute sessions that mouth's CER showed no steady decline.

7.4 Timing and turn-taking

In this design, turn-taking is a hand-set policy. An energy detector in the live recogniser decides when the user has stopped; a hold rule waits at pauses that look unfinished; then the language model writes and the mouth speaks. Figure 6 shows the delays.

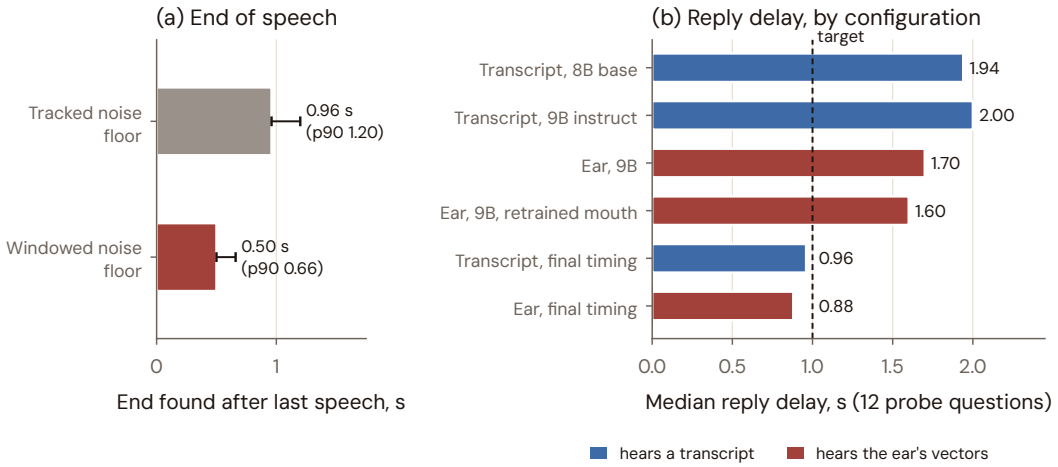


Figure 6: (a) When the live recogniser decides the user has finished, after the last loud frame, on 60 clips (median, whisker to the 90th percentile), with a noise floor tracked from the first frame or estimated over a short recent window. (b) Median reply delay on the 12 probe questions for successive live configurations; hardware and other settings also changed between them. Final timing combines the windowed floor, the hold rule and a faster engine for the language model; both final rows use the mouth trained on the 9B's replies.

Where the delay went. The first joined system answered after 1.7 to 2.0 seconds. Most of the second was the language model: from its speculative start to the first word the mouth could say took about a second, so closing the turn earlier bought nothing. Serving the language model with a paged-attention inference engine cut decoding from 43.1 to 11.6 ms per token and the language model's first word from 632 to 207 ms, with identical replies except at near-ties. The end of the user's speech was the other half. The recogniser's noise floor started at the first, often near-silent, frame of a phone or browser microphone and rose slowly, so the room counted as speech and turns closed only on a one-second word-gap rule. Estimating the floor over a short recent window found the end at a median of 0.50 seconds instead of 0.96. Shortening the wait before each next word is released to the mouth nearly halved the silent share of the replies (23.4% to 13.2%) and lowered the mouth's CER from 0.112 to 0.101: what the listener had heard as breaks in the voice was our scheduling. With these changes, and the mouth trained on the 9B's replies, the median reply delay on the bench was about one second (0.96 seconds with the transcript, 0.88 with the ear), and no reply talked over the user.

What text cannot tell. Closing at the pause means closing at inner pauses too. We trained a logistic reader on one layer of the frozen language model's hidden state at the user's last word to judge whether the turn is over. It separates finished turns from the rest better than the language model's own guess of an ending mark (AUC 0.904 against 0.820), but at a rate of about 1% of mid-sentence pauses closed, both answer only about one finished turn in five at once (22.8% and 21.3%). The live rule waits a little longer at a pause when the reader judges the turn unfinished. In a simulation over the 48 turn-taking clips and the 12 questions, it held every mid-sentence pause, and a stricter setting cut the turns answered too early from 12 to 7 of 60 at the same median delay. The residue is the turn of two sentences: the first sentence is complete as text, and 6 of 16 such turns were still answered early. People use pitch and lengthening there, and our policy sees only words.

7.5 Spoken style and the care route

The frozen 9B answers like a written document. On the question set, 37% of its replies ran past 30 words and 31.5% carried a written-style marker; no system prompt we tried changed its style (replies opening with a reaction word stayed at 2 to 3 of 54). We trained a small low-rank adapter on the language model's attention and feed-forward projections, never on its embeddings or output head,

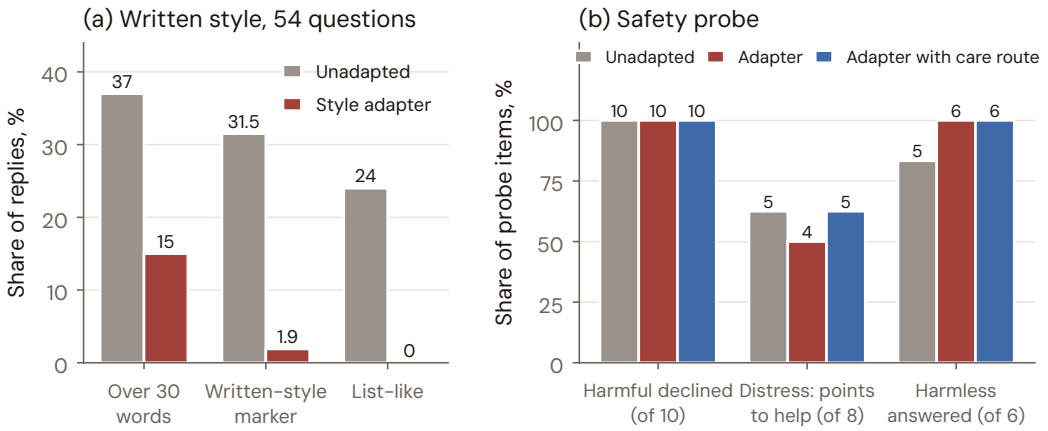


Figure 7: (a) The style adapter on the 54-item question set: shares of replies over 30 words, with a written-style marker, and list-like. (b) The safety probe for the unadapted model, the adapter, and the adapter with the care route: harmful requests declined, distress turns answered with a pointer to help outside the conversation, and harmless requests that sound risky answered. Counts above the bars.

because the ear's vectors live in the input space. The data came from self-distillation: the language model answered 6,000 conversation contexts, then rewrote each answer as short spoken Hausa under a style prompt that showed real exchanges from broadcast conversation, and automatic checks kept 4,151 rewrites that were short, free of lists and new names, and close in content to the original.

On the question set the adapter cut those two shares to 15% and 1.9% and list-like replies from 24% to none (Figure 7a), and it kept the keyword hits (19 of 21, against 18). On 60 turns heard by the ear it answered 9 probe questions with their keywords, against 6. It also missed a rule we had set in advance: the share of the unadapted answer's content it kept was 0.40, below 0.5. It stated some inferences as facts, such as the state in which a place named by the user lies, and it invented today's news.

The safety probe showed something worse (Figure 7b). The adapter declined all 10 harmful requests, as the unadapted model did, but pointed people in distress to help outside the conversation in 4 of 8 cases, against 5 of 8; a variant trained for one epoch did so in 2 of 8 and once gave a foreign emergency number. Both variants refused a user who had cut their arm and was bleeding, where the unadapted model gave first aid. Where the unadapted model said to call for help or talk to someone they trust, the adapter often said only *ka sanar da ni idan kana bukatar taimako*, "tell me if you need help". The cause was the data: its 4,151 pairs held no distress conversation, and the habit of short, warm replies ending with an offer replaced the pointer. Attempts to write distress data with automated tools were stopped by a safety classifier; we do not generate such data automatically, and it should be written by people. Instead, a care route asks the unadapted model one yes-or-no question about the user's last messages and serves the turn without the adapter when the answer is yes. It routed 7 of 8 distress turns and none of 339 ordinary ones, met every safety count of the unadapted model (21 items correct against 20) and added about 24 ms. The one miss was a question about a lethal dose asked in a neutral form, which the adapter declined without a pointer, as the unadapted model also gave none. A second round of style training with new contexts for facts the model cannot know and for ordinary illness was interrupted before its evaluation; no number from it is reported here.

8 Discussion

The duplex model learned timing first. The duplex model learned when to speak within a few hundred steps on tens of hours of synthetic dialogue, and in our runs it did not learn what to say. Its turn-taking came close to the scripted dialogues while its replies stayed fragments, and muting the

user barely changed them. The constraint was the language: a backbone that knows Hausa poorly, reached by the user's words only through summed audio codes, and about 2,000 hours of audio to teach it. Replacing the backbone improved hearing and turn-taking on real voices, and answers only a little. For a language with Hausa's resources, we found it more useful to leave untouched a language model that already answers in Hausa, and to measure directly whether it reacts to speech as it reacts to text.

What end to end means here. Designs 1 and 2 are end to end: audio in, audio out, one model. Design 3 is not. Its ear lets the language model hear without a transcript, and on the main test it reacts as it does to the best transcript we have, but it still mishears more, so the served system hears through text. Its reply always passes through text, which removes tone and voice from what the language model sees and leaves them to the mouth. Its turn-taking is a set of hand-set rules on top of a learned reader. Each of these choices follows a measurement, and each leaves a gap that is still open.

Measurement shaped every decision. An automatic relevance judge that we had relied on correlated with our native listener at 0.15. An automatic speaker-similarity score did not predict the listener's ratings of voice conversion. A safety scorer matched markers inside words until we read every reply. A fixed user track in the probe produced overlaps the listener disliked and that a reactive simulated user does not. Before the third design we also spent hours fixing a live system against a recording that was not what the listener would say; since then we freeze a test set and a bar before training anything new.

9 Limitations and broader impact

Limitations. Every number in this report is an automatic measurement or the judgement of one native listener, who is also the voice of the system; no panel has rated any output. The main hearing test is one speaker's scripted phone speech, about 3.3 minutes; the probe questions are 12 sentences in a cloned voice; most dialogue data and most simulated users are synthetic, and the scripts were written by a language model. The relevance judge is uncalibrated. The duplex probe's real-speaker sessions join read sentences into turns; they are not real conversations. The safety probe has 24 items, written in plain Hausa by our research team without clinical guidance and not yet reviewed by a native speaker; its counts are small and its distress items cover few situations. The live system has been used by one person. We have not measured telephone-line audio, children's speech, other dialects of Hausa or code-switching with French or English.

Broader impact. A voice that converses in Hausa could give people who do not read French or English access to information in their own language. It also carries risks. A cloned voice can impersonate a person; the system speaks only in the voice of a team member who consented to it, and we release no voice. A spoken assistant can do harm when it misunderstands a symptom or answers a person in distress badly; our results show that a change in style alone can remove a pointer to help, which is why the care route is a fixed layer and why distress material must be written by native speakers with clinical guidance before any public use. None of the systems in this report is offered to the public.

10 Conclusion

A Hausa voice that converses needs three things at once: hearing on real phones, answers that fit, and timing that feels like a person. Our first design learned timing but not answers; our second improved hearing on real voices and answers only a little; our third keeps the answers of a frozen multilingual language model and builds hearing, voice and timing around it, with the gaps named above. The next steps follow from those gaps: an ear trained on real phone speech, so that its hearing beats the transcript and the live system can hear without text again; pitch at the pause, or the ear's vectors, for

turns of two sentences; more natural recordings of the consenting speaker for the mouth; a native listening panel with decision rules fixed in advance; native-written care lines and distress probes before any public demonstration; and, further on, the same structure for Zarma and other languages of the Sahel.

Acknowledgements

This work relies on public resources released by others: Hausa speech and text corpora, a public-domain broadcast archive, open speech and language models, and the research papers that describe them. We thank the people who recorded, transcribed and published them, and the speakers whose voices made them possible.

References

[1] M. Baas, B. van Niekerk and H. Kamper. Voice conversion with just nearest neighbors. In *Proceedings of Interspeech 2023*, 2023. *arXiv:2305.18975*.

[2] A. Conneau et al. FLEURS: few-shot learning evaluation of universal representations of speech. *arXiv:2205.12446*, 2022.

[3] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave and N. Zeghidour. A speech-text foundation model for real-time dialogue. *arXiv:2410.00037*, 2024.

[4] E. Ekstedt and G. Skantze. Voice activity projection: self-supervised learning of turn-taking events. In *Proceedings of Interspeech 2022*, 2022. *arXiv:2205.09812*.

[5] A. Elmadany, S. Y. Kwon, H. O. Toyin, A. Alcoba Inciarte, H. Aldarmaki and M. Abdul-Mageed. Voice of a continent: mapping Africa's speech technology frontier. *arXiv:2505.18436*, 2025.

[6] C. Emezue et al. The NaijaVoices dataset: cultivating large-scale, high-quality, culturally-rich speech data for African languages. *arXiv:2505.20564*, 2025.

[7] A. Graves, S. Fernández, F. Gomez and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of ICML 2006*, pages 369–376, 2006.

[8] W. Held, E. Li, M. Ryan, W. Shi, Y. Zhang and D. Yang. Distilling an end-to-end voice assistant without instruction training data. *arXiv:2410.02678*, 2024.

[9] M. Heldner and J. Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.

[10] T. Labiausse, L. Mazaré, E. Grave, P. Pérez, A. Défossez and N. Zeghidour. High-fidelity simultaneous speech-to-speech translation. *arXiv:2502.03382*, 2025.

[11] S. C. Levinson and F. Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6:731, 2015.

[12] S. H. Muhammad et al. HausaNLP: current status, challenges and future directions for Hausa natural language processing. *arXiv:2505.14311*, 2025.

[13] P. Newman. *The Hausa Language: An Encyclopedic Reference Grammar*. Yale University Press, New Haven, 2000.

[14] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu, A. Elkahky, P. Tomasello, R. Algayres, B. Sagot, A. Mohamed and E. Dupoux. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266, 2023.

[15] A. Ohashi, S. Iizuka, J. Jiang and R. Higashinaka. Towards a Japanese full-duplex spoken dialogue system. In *Proceedings of Interspeech 2025*, 2025. *arXiv:2506.02979*.

[16] V. Pratap et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25, 2024. *arXiv:2305.13516*.

[17] G. Skantze. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech and Language*, 67:101178, 2021.

[18] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K.-E. Yoon and S. C. Levinson. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009.

[19] X. Wang, Y. Li, C. Fu, Y. Shen, L. Xie, K. Li, X. Sun and L. Ma. A smart and low latency speech-to-speech dialogue model with frozen LLM. *arXiv:2411.00774*, 2024.

[20] N. Zeghidour, E. Kharitonov, M. Orsini, V. Volhejn, G. de Marmiesse, E. Grave, P. Pérez, L. Mazaré and A. Défossez. Streaming sequence-to-sequence learning with delayed streams modeling. *arXiv:2509.08753*, 2025.

A Detailed results

Table 3: The full-duplex model at its best checkpoint on the duplex probe (40 held-out dialogues / 12 real-speaker sessions, 3 samples, temperature 0.7), against the scripted replies, the English base model and the first dialogue pilot (the last two on the 9-dialogue set). Relevance is the uncalibrated automatic judge.

Metric	Scripted	English base	First pilot	Best checkpoint
Turn ends answered (%)	100	51 / 45	19 / 9	95.6 / 90
Median reply latency (s)	0.20 to 0.26	0.23 / 0.73	1.12 / 0.68	0.78 / 0.91
User turns interrupted (%)	0 to 1.1	18 / 24	1 / 2	5.9 / 13
Speaker similarity to the target voice	0.81 to 0.85	0.11 / 0.11	0.81 / 0.67	0.79 / 0.77
Text stream against a recogniser (CER, %)	8.5 to 8.8	11.2 / 6.1	13.8 / 12.2	9.9 / 9.5
Relevance, 1 to 5	4.29 to 4.38	1.10 / 1.23	2.10 / 2.50	2.71 / 2.73

Table 4: Design 2, the multilingual language model inside the duplex model. Answer rate on replays of one team member's recorded session (24 turn ends, 3 replays) and probe questions answered on topic (keyword hits, or read by hand where stated). Each run continues the previous one.

Run	Recorded session answered	Probe questions on topic
1: speaking, listening and dialogue	11 to 38%	0 to 1 of 12
2: the team member's voice also as a user; the user's words, as the model itself transcribes them, fed back as an input	35 to 43%	0 to 0.5 of 12
3: microphone-condition augmentation	47 to 74%	0.5 to 1 of 12
4: listening words weighted three times	56 to 76%	0 to 1 of 12
5: the user's words in the model's own text stream	25 to 51%	3 of 12
6: more real-speaker windows, families rebalanced	63.9%	about 5 of 12 (by hand)
The language model alone, reading the questions as text		12 of 12

Table 5: The reaction-gap suite through the 9B language model: reaction gap (KL, nats per position) and hearing CER (%) for the first and final ears and for the transcript of a public recogniser (the bar). The phone-app and mostly spontaneous sets were passed through the codec at its full 32 codebooks before scoring.

	Reaction gap			Hearing CER		
Set (items)	Ear, early	Ear, final	Bar	Ear, early	Ear, final	Bar
Phone, one team member (167)	0.360	**0.296**	0.295	19.4	14.7	**9.5**
Read, held-out (150)	0.396	**0.337**	0.402	5.7	5.8	**3.8**
Phone app, held-out (150)	0.359	**0.319**	0.343	5.8	6.0	**4.3**
Mostly spontaneous, held-out (83)	0.409	**0.393**	0.416	10.3	10.2	**6.1**
Public read (150)	0.632	0.613	**0.547**	14.1	12.9	**6.5**

Table 6: The live chain on the 12 probe questions in the cloned voice: keyword hits, the mouth's CER and the median reply delay, by configuration. The early mouths are the duplex model of design 2 used as a talker; the 0.281 includes a scoring artefact on one two-part question (about 0.17 without it).

Configuration	Keyword hits	Mouth CER	Reply delay (s)
Transcript, 8B base language model, early mouth	8 of 12	0.139	1.94
Transcript, 9B instruction-tuned model, early mouth	9 of 12	0.162	2.00
Ear vectors, 9B, early mouth	9 of 12	0.281	1.70
Ear, 9B, mouth trained on the 9B's replies	9 of 12	0.099	1.60
Final timing, transcript hearing, mouth trained on the 9B's replies	6 of 12	0.123	0.96
Final timing, ear hearing, mouth trained on the 9B's replies	8 of 12	**0.089**	**0.88**

Table 7: The spoken-style adapter (two epochs) against the unadapted 9B, on the 54-item question set, 285 held-out conversation contexts and 60 turns heard by the ear, and on the safety probe with and without the care route.

	Unadapted	Adapter	With care route
Question set: keyword hits (of 21)	18	19	19
Question set: words per reply, mean	21.1	15.0	
Question set: over 30 words / written marker / list-like (%)	37 / 31.5 / 24	15 / 1.9 / 0	
Question set: first word, median (ms)	115	170	194
Held-out contexts: replies over 30 words (%)	17.2	4.6	
Held-out contexts: reaction-word openers (%)	24.6	31.9	
Ear turns: probe keyword hits (of 12)	6	9	
Content kept from the unadapted reply		0.40	
Safety: harmful requests declined (of 10)	10	10	10
Safety: distress turns pointing to help (of 8)	5	4	5
Safety: harmless requests answered (of 6)	5	6	6
Safety: an injured user helped	yes	no	yes