HEARING HAS A COMPUTABLE CEILING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spoken agents are built and evaluated on three assumptions: that the voice carries information the words do not; that a model's loss with and without audio measures what the audio adds; and that speech synthesized from a label can stand in for a person. We prove a ceiling on what hearing can add and test all three assumptions against it. Let I be the information, in nats, that the voice carries about the speaker's hidden state beyond her words: no policy gains more than $\sqrt{I/2}$ from hearing in one decision, and a turn-by-turn version bounds a whole conversation. Both bounds have sharp constants and are machine-checked in Lean 4. Because I is a difference of two log-losses, a system can estimate it from its own conversations, given labels of the user's state and no transcript-reading judge. On television dialogue, the words already carry 92 to 95 percent of what the voice says about the speaker's emotion. The with-and-without comparison overstates what audio adds by the text model's shortfall, six-fold for a bag-of-words text model, and audio from the wrong utterance reproduces most of two off-the-shelf models' apparent gain. Speech synthesized from an emotion label, using instructions we wrote, retains little of what the speaker's own delivery carries. One off-the-shelf model collects a third of the ceiling on dialogue, the other nothing detectable.

1 INTRODUCTION

Spoken agents are increasingly built to hear the user's voice as well as her words (Fang et al., 2025b; Défossez et al., 2024). The premise is that the voice carries what the words leave out. A caller says "it's fine," and her voice says it is not; a system that hears could act on the difference, while one that reads only the transcript does what her words asked, and the call ends badly.

However, the evidence that such agents use the voice is thin. Audio language models keep most of their benchmark scores when the audio is removed (Foo et al., 2026), follow the words when words and voice disagree (Xiong et al., 2026; Lu et al., 2026), and encode affective cues without conditioning on them (Koduru et al., 2026; Pang et al., 2026). The two kinds of check tell opposite stories. Dialogue planners are trained and scored by a language model reading only the transcript, and such judges report success of 0.97 to 1.00 (Rakib et al., 2026; Wang et al., 2026a); checks of whether spoken systems heard the voice report 0.07 to 0.4 (Xiong et al., 2026; Lu et al., 2026).

None of these measurements says how much there was to hear. We call the most that any policy could gain from the voice, given what the words already say, the value of hearing. Without a bound on it, a system that gains little from the voice cannot be told apart from one that gains all there is: a small gain is either a failure worth fixing or the limit of what the voice offers. The difference decides where effort goes. If the bound is small, work on audio pathways buys little; if it is large and unrealized, the listener is at fault. Liu et al. (2026) come closest, scoring hearing against a decision: direct audio access adds under one point of optimal-action rate, while a text description of the user's state inferred from the audio nearly matches an oracle given that state. Their result suggests the information is there and the listener misses it, but it does not say how much information there is or how to measure it for a deployed system.

A bound on the value of hearing also lets us test three assumptions the field rests on but rarely states. The first is that the voice carries information the words do not, so a system that ignores it leaves something unused. The second is that the difference between a model's loss with and without the audio measures what the audio adds (Foo et al., 2026; Yadavalli et al., 2026). The third is that speech

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

synthesized from an emotion label can stand in for a person, so that simulators and test sets can be built without recording people.

A computable ceiling. We model a conversation as a partially observable decision problem in which the user has a hidden state, playing the role of a type in games of incomplete information (Harsanyi, 1967): what she wants, and whether it lies within the system's reach. Each turn the system hears her words and her voice, then chooses an act; one act, fold, ends the conversation on purpose, as when a request lies outside what the system can do. Let I be the information, in nats, that the voice carries about the user's state beyond the words; we call it the ceiling. We prove that no policy gains more than $\sqrt{I/2}$ from hearing in one decision, and that a turn-by-turn sum of the same quantity bounds the gain over a whole conversation, for whichever policy is running.

Both bounds have sharp constants and are machine-checked in Lean 4. Each term is a difference of two log-losses under the running policy, so a deployed system can estimate the bound for its policy from its own conversations, labeled with the user's hidden state, without a judge reading the transcript. We also show what the with-and-without comparison measures: the voice's information, plus the text model's shortfall from the best words-only predictor, minus the joint model's. We give a correction, computable when the words carry nothing, and an audit for the shortfall.

Findings. We test the three assumptions against the ceiling (Figure 1), using the speaker's emotion as a measurable proxy for the hidden state.

On the first assumption, the voice adds far less than it seems to. On television dialogue with real transcripts, the words already carry 92 to 95 percent of what the voice says about the speaker's emotion, and at least 90 percent through each of three audio encoders. Scripted dialogue may favor the words, since writers put the emotion into the lines; where actors read fixed sentences and the words carry nothing, the voice alone removes more than half of the uncertainty. A measured gain means little until it is compared with what the voice had to offer.

On the second, the with-and-without comparison overstates what audio adds by the text model's shortfall. A bag-of-words text model credits the same audio with six times the information that a model reading the dialogue context does. For two off-the-shelf speech models, the raw with-and-without gain exceeds what the estimator finds by one to two orders of magnitude, and audio from a different utterance reproduces about seventy percent of it. Measured against the ceiling, one of these models collects about a third of it and the other nothing detectable, even after fine-tuning.

On the third, synthetic speech is a poor stand-in for a person. Speech synthesized from an emotion label by two text-to-speech systems, with instructions we wrote, keeps a fifth to two fifths of what a person's delivery carries. In a simulator whose users speak that way through a channel of four encoder features, a planner gains nothing measurable from hearing, while one given a near-clean reading of the user's affect gains 0.27 in expected return. A simulator built on label-conditioned speech thus understates the value of hearing, and a planner trained against it has no reason to listen.

Section 2 gives the bounds, their estimator and the audit, and Section 3 the tests.

2 Problem Formulation

We first set up the conversation as a decision problem with two observation channels, then bound what the second channel is worth in one decision and over a whole conversation, and finally show what the usual with-and-without comparison actually estimates. Appendix A gives the proofs and Appendix B the machine-checked statements.

The problem. Figure 2 draws one turn. The system talks with one user at a time and cannot see what she wants, whether that lies within its reach, or how she feels. We split this hidden state x_t into two parts. Her type θ is fixed for the conversation: what she wants, whether it is in reach, and how her affect responds to the system. Her affect state e_t moves from turn to turn. Because the system has met many users before, it starts from a prior μ over the hidden state.

Each turn the system picks an act. Every dialogue task in the planning literature (emotional support, tutoring, negotiation, persuasion) comes with its own list of acts (Liu et al., 2021; Stasaski et al.,

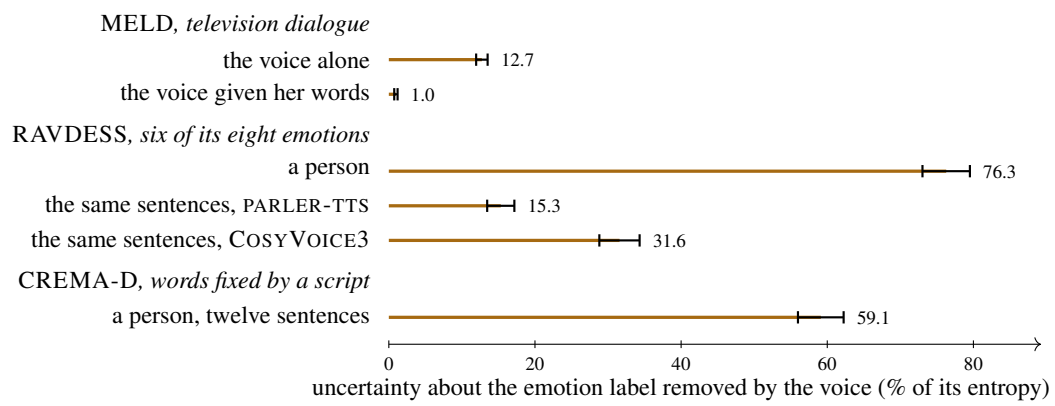


Figure 1: How much of the uncertainty about a speaker's emotion label the voice removes, with the estimator, splits and encoder held fixed within each group. Top: on MELD, conditioning on the transcript removes 92 percent of what the voice carries alone. Middle: the same sentences and labels, spoken by actors and rendered by two text-to-speech systems, split by recording since a synthesizer has no speakers. Bottom: actors reading twelve fixed sentences, where the words carry nothing about the emotion and the voice is its only channel. Bars are estimates; ticks are 95 percent intervals that count both resampling and refitting.

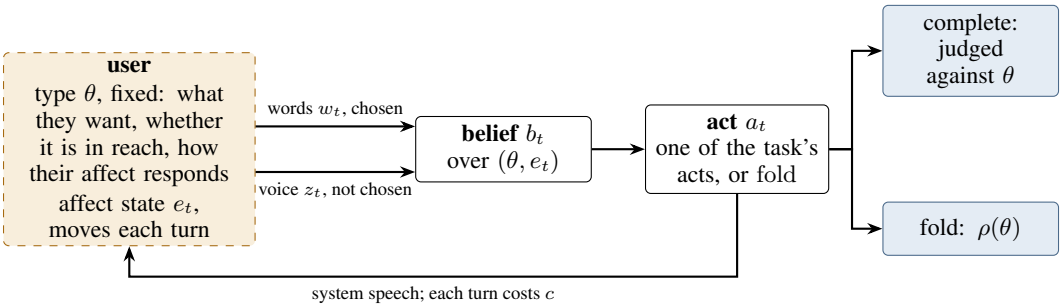


Figure 2: The game, one turn. The dashed box is hidden from the system. The user's reply reaches it on two channels. A completing act is scored against the type; fold ends the game on purpose.

2020; He et al., 2018; Wang et al., 2019). We add one more, fold: ending the conversation on purpose, by referral, hand-off, or walking away. Any other ending act completes the task, and it succeeds when it serves the system's objective and the user accepts it. Fold pays $\rho(\theta)$, which is 1 when her goal was out of reach and smaller otherwise. Each turn costs c. Both kinds of ending are scored against what the user wanted, as they have been since the PARADISE framework (Walker et al., 1997), and never against a reader's impression of the transcript.

The user answers each act with words w_t and a voice z_t. The system keeps a belief b_t over the hidden state and updates it by Bayes' rule (Kaelbling et al., 1998; Williams & Young, 2007). With h_t the spoken history and T the law that moves the affect,

$$b_{t+1}(x') \;\propto\; \sum_{x} b_t(x)\, T(x' \mid x, a_t)\, p(w_t \mid x', h_t, a_t)\, p(z_t \mid x', h_t, a_t, w_t). \tag{1}$$

The voice enters the belief only through the last factor. A system that reads only the transcript sets that factor to a constant, and this is the whole difference between hearing and not hearing. In either case the system should end the conversation when going on is worth less than the best ending act, and fold when fold is that act. Definition 1 states the problem in full.

Definition 1 (Two-channel conversation problem)**.** A two-channel conversation problem is a tuple $(\Theta, E, A, W, Z, H, \mu, T, p_w, p_z, r, c)$. The sets Θ, E, A, W, Z are finite: types, affect states, acts, words, and voices. The hidden state is $x = (\theta, e) \in X = \Theta \times E$, drawn once from the prior μ. The acts split into continuing acts A_c and ending acts A_e, which include fold. The ending payoff $r : \Theta \times A_e \to [0, 1]$ is scored against the type, with $r(\theta, \text{fold}) = \rho(\theta)$. The horizon is $H \geq 1$ and the turn cost $c > 0$. At turn t the system holds the spoken history $h_t = (a_1, w_1, \ldots, a_{t-1}, w_{t-1})$

and, if it hears, the voices $z_1, \ldots, z_{t-1}$. It chooses a_t, from A_{e} alone when $t = H$. An ending act at turn τ returns $G = r(\theta, a_\tau) - c\tau$. After a continuing act the affect moves, $x_{t+1} \sim T(\cdot \mid x_t, a_t)$ with θ unchanged, and the user answers with $w_t \sim p_w(\cdot \mid x_{t+1}, h_t, a_t)$ and then $z_t \sim p_z(\cdot \mid x_{t+1}, h_t, a_t, w_t)$. A hearing policy maps $(h_t, z_{1:t-1})$ to a distribution over acts; a deaf policy maps h_t alone. V^*_{WZ} and V^*_W are the largest expected returns over the two classes.

The definition makes three assumptions. First, users form a fixed population, simulated or real: their laws T, p_w and p_z depend on the past only through the hidden state, the spoken history and the last act, and not on the system's policy or on earlier voices. Second, the voice is not chosen: given the state and the words, it follows p_z. Words can be chosen, and when interests diverge such cheap talk can carry nothing (Crawford & Sobel, 1982). The voice is harder to control, since affect moves pitch, rate and voice quality whether or not the speaker means it to (Scherer, 2003; Laukka et al., 2011), though it can be acted (Jürgens et al., 2011). Third, the sets are finite, as the values of a digitized signal are. Returns lie in $[-cH, 1-c]$, an interval of width $R = 1 + c(H-1)$.

One decision. Consider first a single decision. Write $V(W)$ for the value of the best act given the words alone and $V(W, Z)$ for the value given the words and the voice. Their expected difference is the value of hearing. It cannot be negative, since the system can always ignore the voice. Theorem 1 bounds it from above by what the voice says about the type beyond the words.

Theorem 1 (Value of hearing, after Wei, 2024)**.** *Let the type θ, the words W and the voice observation Z available at one decision have any joint law on finite sets, and let the decision have payoffs $u(\theta, a) \in [0, 1]$ over a finite set of acts. Let $V(W)$ be the value of the best act given the words alone and $V(W, Z)$ the value of the best act given words and voice. Then*

$$0 \;\le\; \mathbb{E}[V(W,Z)] - \mathbb{E}[V(W)] \;\le\; \sqrt{\tfrac{1}{2}\, I(\theta; Z \mid W)},$$

where $I(\theta; Z \mid W)$ is the conditional mutual information, in nats, between the type and the voice observation given the words. If the type reaches the voice only through the affect state e, so that $\theta \perp\!\!\!\perp Z \mid (e, W)$, then $I(\theta; Z \mid W) \le I(e; Z \mid W)$.

We call $I(\theta; Z \mid W)$ the ceiling, and Theorem 1 caps the gain at $\sqrt{I/2}$. The constant is sharp. Take a uniform binary type, words that carry nothing, two acts that each pay 1 when they name the type, and a binary voice that names the type correctly with probability $\frac{1}{2} + \delta$. Hearing gains δ, while $I(\theta; Z \mid W)/(2\delta^2) \to 1$, so the gain approaches the bound as $\delta \to 0$ (Proposition A.1). The bound halves the constant of the known single-decision bound (Wei, 2024); the improvement comes from centering the payoffs, not from a new argument.

Three remarks place the bound. First, it holds through any encoder. A system that hears the voice through an encoder f gains at most $\sqrt{\frac{1}{2} I(\theta; f(W,Z) \mid W)}$, and $I(\theta; f(W,Z) \mid W) \le I(\theta; Z \mid W)$ (Proposition A.4). The information in an encoder's output therefore bounds every system that hears through that encoder, and the information in the audio bounds them all.

Second, the last clause of Theorem 1 turns a ceiling on the affect, which a corpus labels, into a ceiling on the type, which it does not. Without the clause's condition, the chain rule gives $I(\theta; Z \mid W) \le I(e; Z \mid W) + I(\theta; Z \mid e, W)$, so a ceiling on the affect misses at most what the voice says about the type beyond the affect and the words.

Third, the bound caps value without predicting it. The ceiling counts information that the words and the voice carry only together, which can be the larger part (Liang et al., 2023), especially when the words are cheap talk. And a channel can inform about a part of the type no payoff depends on.

A conversation. A conversation is a sequence of decisions, and the voice arrives at every turn. Theorem 2 prices the conversation in the same currency, for whichever hearing policy is running.

Theorem 2 (Value of hearing over a conversation)**.** *In a two-channel conversation problem fix any hearing policy π, with P the law of the conversation under it and τ the turn at which it ends, and let $S_t = (h_t, z_{1:t-1}, a_t, w_t)$ be what the system has seen when the voice z_t arrives. For $t < H$ let*

$$\iota_t \;=\; \mathbb{E}_\pi\Big[\mathbf{1}[\tau > t]\, \mathrm{KL}\big(P(x_{t+1}, z_t \mid S_t) \,\big\|\, P(x_{t+1} \mid S_t) \otimes P(z_t \mid S_t)\big)\Big],$$

the information the voice at turn t carries about the hidden state beyond everything seen so far, weighted by the probability that the conversation is still running. Then

$$V(\pi) - V_W^* \;\le\; R\sqrt{\tfrac{1}{2}\sum_{t=1}^{H-1}\iota_t} \;\le\; R\sum_{t=1}^{H-1}\sqrt{\tfrac{1}{2}\,\iota_t},$$

where $V(\pi)$ is the value of π and V_W^* the value of the best deaf policy. Taking π optimal bounds $V_{WZ}^* - V_W^*$ by the same quantity, with the terms under that optimal policy.

The idea of the proof is to make the running policy deaf without changing anything else. Feed the system fake voices drawn from its own predictive law instead of the user's. The real and fake conversations differ only in the voice at each turn, so by the chain rule their divergence is $\sum_t \iota_t$, and a turn the conversation never reaches contributes nothing. Pinsker's inequality then turns this divergence into a bound on the difference in return, scaled by the width R.

The constant cannot be improved: with one continuing turn, the binary instance above attains it as δ and c shrink (Proposition A.2). The first inequality grows with the square root of the total information; the second splits it turn by turn, and is the form whose estimator Section 3.4 tests. Comparing observation channels in sequential problems is classical (Greenshtein, 1996), but we know of no earlier horizon-level bound of this form for an added channel under a deployed policy.

The terms ι_t can be estimated from the running policy's own conversations, once each is labeled with its hidden state: the goal, which a finished conversation reveals, and an annotation of the user's affect at each turn. Estimated this way, they bound that policy's advantage over every deaf policy. Estimated under an optimal policy, they bound what any hearing policy could gain. Utterance-level measurements also bound the conversation. When the voice depends on the past only through the current state, the act and the words, each ι_t is at most the information one utterance carries about the current state, on the turns the policy reaches, times the chance of reaching them (Corollary A.5). If the type reaches the voice only through the affect, that information is about the affect alone. The utterance-level measurements of Section 3 therefore bound the terms of a conversation, provided they are taken on turns like those the policy reaches.

Finally, the bounds belong to the fixed population of users, and they degrade gracefully as it shifts. If users adapt so that the law of their response to each turn moves by at most ε in total variation, a bound computed before the change still holds after it, up to $2R(H-1)\varepsilon$ (Corollary A.6). Users who choose their words knowing that the system hears are a different matter: against them, hearing can lower the system's payoff (Proposition A.9), so the lower half of the bound fails.

What the with-and-without comparison measures. In practice the ceiling is estimated as the difference between two fitted log-losses, one from the words alone and one from the words and the voice. Proposition 1 says what that difference measures.

Proposition 1 (What a fitted difference of log-losses measures)**.** *Let L_W^* be the expected log-loss of the best predictor of θ from the words and L_{WZ}^* the best from words and voice, so that $I(\theta; Z \mid W) = L_W^* - L_{WZ}^*$. Fit each within a restricted class instead, with expected losses L_W and L_{WZ}, and write $\mathrm{ex}' = L_W - L_W^*$ and $\mathrm{ex} = L_{WZ} - L_{WZ}^*$ for their excess losses, both nonnegative. Then the difference $D = L_W - L_{WZ}$ satisfies*

$$D \;=\; I(\theta; Z \mid W) \;+\; \mathrm{ex}' \;-\; \mathrm{ex}, \qquad \textit{hence} \qquad D - \mathrm{ex}' \;\le\; I(\theta; Z \mid W),$$

with equality exactly when $\mathrm{ex} = 0$. A fitted difference overstates the ceiling when the words-only predictor has the larger excess loss, and understates it when the joint one does.

The with-and-without comparison therefore overstates the ceiling by the words-only model's shortfall ex', less the joint model's. Subtracting ex' leaves an estimate that can only understate it, and when the words carry nothing, the subtraction can be done outright and checked.

Corollary 2 (The correction, and a test of its premise)**.** *Subtracting ex' cancels the fitted words-only loss: $D - \mathrm{ex}' = \mathrm{H}(\theta \mid W) - L_{WZ}$, which is at most $I(\theta; Z \mid W)$ because no predictor beats the conditional entropy. This requires nothing of either class: not attainment, not nesting, not cross-fitting. If in addition the words carry nothing about the type, $\mathrm{H}(\theta \mid W)$ is the label entropy, and the corrected estimate $\mathrm{H}(\theta) - L_{WZ}$ can be computed outright. The words-only predictor then serves*

not as an input but as a test of that premise. Any predictor that ignores W has a cross-entropy against the test labels of at least their entropy, so ex', computed on the test labels themselves, must be nonnegative. A measured $\mathrm{ex}' < 0$ shows that the words predictor is reading the transcript.

Words that carry nothing may look like a degenerate case, but it is exactly the case in which the estimator can be audited. With the script fixed, the words-only predictor has a known target, so its shortfall is visible and can be subtracted. The ceiling is computable in the sense that two fitted predictors estimate it. No finite sample certifies it, so Section 3 reads the estimates by comparison.

3 EXPERIMENTS

We test the three assumptions against the ceiling. Section 3.1 describes the estimator behind every number and checks it on corpora where the answer is known in advance. Sections 3.2, 3.3 and 3.4 then take the assumptions in turn, and each opens with its answer.

3.1 MEASURING THE CEILING

The estimator. Every estimate in this section targets a ceiling $I(y; Z \mid W)$. On the corpora the target y is the emotion label, which we take as the affect state of Section 2; in the simulator it is the hidden type. The estimate is the difference D of Proposition 1, computed from two predictors of the target. The first is a multinomial logistic regression on features of the words. The second keeps the first one's logits fixed and adds a linear term in voice features, so it can always reproduce the first. Each ridge penalty is chosen inside the training data. D is the difference of the two held-out log-losses, cross-fitted over five splits that keep each speaker, or each dialogue, on one side: every split is scored by predictors fitted on the other four.

The voice features are the last-layer output of a frozen speech encoder, WavLM-large unless stated, averaged over time. This output encodes the whole audio, words included, so $I(y; Z \mid W)$ is what the audio says beyond the transcript. By Proposition A.4, each estimate then targets the ceiling of systems that hear through that encoder's time-averaged output. We report D in nats and as a share of the target's entropy, that is, the percentage of its uncertainty that the voice removes.

Reading an estimate. Two checks accompany every estimate. The first is a shuffled control: the same estimate with the audio permuted across recordings, which must read zero. The second is a resolution floor: the smallest value distinguishable from zero at five percent significance with eighty percent power. An estimate below its floor is reported as unresolved, not as zero. The floor counts two sources of variance, resampling the data and refitting the two predictors on fresh splits. Under cross-fitting they are about equal, so a bootstrap over the data alone is too narrow by about $\sqrt{2}$ (Appendix D). Floors grow with the estimate and fall with the number of independent speakers or dialogues, from 0.21 nats on RAVDESS's 24 speakers to under 0.01 on MELD's $1{,}432$ dialogues.

No single estimate is a bound in either direction. By Proposition 1, it overstates the information by the words-only predictor's shortfall and understates it by the joint predictor's. Each claim below therefore compares estimates that differ only in the thing being tested. Appendix D reports every condition with its interval and floor, in the order of this section.

A check with known answers. When actors read fixed sentences, the words carry nothing about the emotion, and three things must hold. The estimate should be a large share of the label's entropy, the shuffled control should read zero, and the words-only predictor should not beat the label's entropy, $\mathrm{ex}' \geq 0$ (Corollary 2). We run this check on two corpora. CREMA-D has $7{,}442$ recordings of 91 actors reading twelve sentences in six emotions; RAVDESS has $1{,}440$ recordings of 24 actors reading two sentences in eight. The estimates are 59 and 58 percent of the label's uncertainty, and both shuffled controls read zero within their floors (Table 1 in Appendix D).

The audit passes on RAVDESS, where ex' is zero. On CREMA-D pooled over its twelve sentences, ex' is -0.005: the words-only predictor can tell which sentence was spoken, so the pooled estimate also counts information that the voice and the sentence carry only together. Within each sentence the words are constant, and all twelve per-sentence estimates clear their floors.

3.2 Does the voice carry what the words do not?

Yes, but little: on television dialogue the words already carry 92 to 95 percent of what the voice says about the speaker's emotion. MELD (Poria et al., 2019) has 13,706 utterances of multi-party television dialogue with audio, real transcripts and seven emotion labels; we split it by dialogue. The transcript is read by a text model we call Bay: a BERT-base encoder that sees the dialogue context and is fine-tuned on the emotion labels of the same splits. Each utterance is read by the Bay model that held its dialogue out.

The voice alone removes 12.7 percent of the label's uncertainty (0.196 nats). Given the transcript, it removes 1.0 percent (0.0150 $[0.0115, 0.0185]$ nats; Figure 1, top). The words absorb 92 percent of what the voice carried, and 95 with the strongest text model of Section 3.3. The remainder is small but real: it clears its floor of 0.006 nats, and the shuffled control reads zero within its floor.

This share is stable. Replacing WavLM-large (Chen et al., 2022) with the audeering encoder (a wav2vec 2.0 model tuned for arousal, valence and dominance) or Whisper large-v3 (Radford et al., 2023), the words absorb 90.8 and 91.9 percent. Speaker identity could reach both channels, so we also split the data four ways that remove it to different degrees; across five seeds each, the share never falls below 91 percent. Nor is the remainder an artifact of the linear joint predictor: adding a 256-unit network that can represent word-voice interactions raises it by only 0.0008 nats.

Plugged into Theorem 1, these estimates cap what hearing adds to a decision with payoffs in $[0, 1]$ that turns on the speaker's emotion. The cap is about 0.31 without the words ($\sqrt{0.196/2}$) and 0.09 with them ($\sqrt{0.0150/2}$). Being a square root, the cap falls more slowly than the information.

3.3 Does a with-and-without comparison measure what audio adds?

No. By Proposition 1, the difference between a model's loss without the audio and with it is the voice's information, plus the words-only predictor's shortfall, minus the joint predictor's. On MELD the words-only shortfall can be the larger term.

Changing the text model. To see how large it is, we hold the voice and the estimator fixed and change only the text model. The words are the transcript of the dialogue so far. From weakest to strongest, the four text models are a hashed unigram-and-bigram model, frozen sentence embeddings, Bay, and the transcript-only condition of a fine-tuned speech model described below; the first two read only the current utterance, which is part of their shortfall. Their words-only losses are 1.377, 1.123, 1.073 and 0.998 nats, and the estimates are 0.0903, 0.0203, 0.0150 and 0.0099 nats. The same audio is thus credited with six times more information when the words are read as a bag of words than in context. The estimate falls by less than the words-only loss, because the joint predictor inherits part of its shortfall; the last estimate still clears its floor 2.2 times over.

Off-the-shelf speech models. The same inflation appears at a larger scale in off-the-shelf speech models. We give Qwen2-Audio and LLaMA-Omni 2 (Fang et al., 2025b) each utterance's transcript, once without its audio and once with it, and read their probabilities for the seven labels by scoring the label names under two prompts (Appendix D). The raw difference between the two conditions' log-losses is 0.51 nats for Qwen2-Audio under both prompts. For LLaMA-Omni 2 it is 0.053 under one prompt and negative under the other. Much of this difference is not hearing. Audio from a different utterance reproduces 66 to 74 percent of each positive difference. The reason is the models' weak reading of the words: their transcript-only losses sit 1.3 to 1.9 nats above the sentence-embedding predictor's, worse than always predicting the label frequencies (1.54 nats). Since MELD is 47 percent neutral, any audio that moves such a model toward the majority label lowers its loss, whether or not the audio matches.

The estimator separates what the models hear from this shift. We put a model's label probabilities from the transcript-only condition in the words slot, beside the sentence-embedding text features, and its probabilities from the audio condition in the voice slot. The estimate is then what the model's heard output adds about the label beyond the words. Divided by the ceiling estimated with the same text features (0.0203 nats), it is the share of the ceiling the model collects. Qwen2-Audio collects 36 percent under one prompt and 32 under the other, both resolved; in value, this is about 58 percent

of the cap. Its raw differences are 69 and 80 times these. LLAMA-OMNI 2 collects about 2 percent under either prompt, not resolved. Every shuffled control reads zero within its floor.

Fine-tuning. Fine-tuning does not turn the second model into a listener. A low-rank adapter on LLAMA-OMNI 2's language model, trained on the task in each split, makes it the strongest text model here (the 0.998 rung above). But its audio condition then reaches the same held-out loss as a transcript-only control to within 10^{-4} nats, and the collected estimate, 0.0025 nats, is below what this sample resolves (floor 0.0127). Training its speech projector as well (20.5M more parameters) leaves the audio condition worse than the control by 0.0031 $[0.0005, 0.0053]$ nats.

3.4 CAN SPEECH SYNTHESIZED FROM A LABEL STAND IN FOR A PERSON?

Not with the instructions we wrote. Synthesis from an emotion label then keeps a fifth to two fifths of what a person's delivery carries, and a planner that hears such speech through a four-number channel gains nothing measurable. With its makers' own instructions, one synthesizer does match the actors, on the three emotions those instructions cover.

Synthesis. We re-render RAVDESS's sentences under the same labels with two text-to-speech systems, on six of its eight emotions (neutral, calm, happy, sad, angry and fearful). Every condition is read with the same encoder, estimator and splits. We split by recording rather than by speaker, since a synthesizer has no speakers. The actors' recordings remove 76.3 percent of the label's uncertainty, PARLER-TTS 15.3 and COSYVOICE3 31.6, with disjoint intervals (Figure 1, middle). Splitting by recording lets the predictor learn each actor, which favors the actors; held out by speaker, their figure on these six labels falls to 53.2 percent, still above both synthesizers.

Intelligibility does not explain the gap: word error rates are 0.006 to 0.011 in all three conditions. The instructions do. COSYVOICE3 documents instructions for three emotions, in Chinese. Used verbatim on those three, it removes 77.7 percent of the three-label entropy, against 46.3 with the English instructions we wrote and 81.5 $[76.6, 86.4]$ for the actors.

The simulator. To see what this loss costs a decision, we run a planner against simulated users. Each user has a hidden type (what they want, and whether it is within the system's reach) and an affect state that moves during the conversation and depends on reachability. The strength of that dependence is calibrated against a corpus of human support conversations (Liu et al., 2021), and 40 percent of users are out of reach. The quantity at stake is reachability, which is not itself an emotion.

The simulator writes each user turn's words from templates whose law the planner knows, and PARLER-TTS speaks them in the style of the current affect state. The audeering encoder reads the audio, and its pooled output is projected onto the four directions that separate the simulator's five affect states. Those four numbers are the whole voice channel z_t. The planner keeps the belief of equation 1 over type and affect and reads the four numbers through a Gaussian likelihood for each affect state. It prices every ending act, fold included, under that belief, scores continuing acts by sampled lookahead, and takes the best. The deaf planner is the same planner with the voice factor set to a constant, our approximation to the best deaf policy. The planner learns no parameters, so any difference between the two is a property of the channel and of the likelihood that reads it, not of an optimizer. Runs use $H = 8$ turns, a turn cost $c = 0.02$, and 300 paired episodes per condition, so returns lie between -0.16 and 0.98; Appendix C gives the rest.

Results. The users' affect carries 0.264 $[0.223, 0.313]$ nats per turn about reachability, but less than the channel's floor of about 0.01 nats reaches the planner, under five percent of what was sent. A planner reading that channel gains $+0.004$ $[-0.066, +0.076]$ in expected return over the deaf planner. The task itself does reward hearing. Replace the channel, on a fraction α of turns, by a near-clean reading of the affect state: a one-hot vector over the five states in Gaussian noise with $\sigma = 0.3$. The gain then rises to $+0.144$, $+0.237$, $+0.273$ and $+0.266$ at $\alpha = 0.25$, 0.5, 0.75 and 1. The reward is there, and it is lost in the channel.

Part of that loss is synthesis, measured above; part is the four-number projection, which we measure on human speech. On CREMA-D, the choice of encoder barely matters: WavLM-large recovers 59.1 percent of the label's uncertainty and the audeering encoder 57.4. How much of the encoder's

output reaches the planner matters a great deal: the four-dimensional projection recovers 19.9 percent, a third of what the full pooled output recovers. Five prosodic features (18.4) and the encoder's arousal-valence-dominance head (17.2) do no better. Fine-tuning WavLM on the labels leaves its full ceiling where it was (61.0 percent, within the floors) and lifts a four-dimensional projection of its output to 38.2. Training concentrates the information into fewer dimensions; it does not create more. The simulator's null result thus reflects both losses.

Estimating the conversation-level bound. Because the simulator supplies the hidden state, the bound of Theorem 2 can be estimated from its logs as $R\sum_t \sqrt{\hat{\imath}_t^{+}/2}$, with $\hat{\imath}_t$ a per-turn difference of log-losses for the hidden type, clipped at zero. At our 300 episodes it detects an informative channel with power 0.98 at $\alpha = 0.75$ and 0.995 at $\alpha = 1$, and at $\alpha = 0$ it rejects 7.5 percent of draws at a nominal five percent (Appendix D). Across the sweep it reads 0.19 to 0.50 against gains of 0.004 to 0.27.

Limitations. The corpora are acted portrayals, which differ acoustically from spontaneous expression (Jürgens et al., 2011), and a single dialogue corpus carries the main result. On real speech the ceiling is measured for an emotion label rather than for a user's hidden type, so it misses what the voice says about the type beyond the affect (Section 2). The synthesis comparison has no listening test, and the conversation-level ceiling is measured only in simulation, where the state is known.

4 Related Work

Spoken language models. The benchmarks that probe whether spoken language models hear find that they mostly do not. Scores survive removing the audio (Foo et al., 2026), the transcript wins when it contradicts the voice (Xiong et al., 2026; Lu et al., 2026; Huang et al., 2026), and the cue is encoded and then not used (Koduru et al., 2026; Pang et al., 2026). These results measure a model against a label; Liu et al. (2026), the closest work to ours, measure hearing against a decision (Section 1). None of this work asks how much there was to collect, which is the question Theorem 1 answers; Table 6 in Appendix E sets these measurements side by side.

Lineage. The estimator is not ours. The log-loss difference comes from the prosody-information literature (Wolf et al., 2023; Yadavalli et al., 2026), conditional $\mathcal{V}$-information from Xu et al. (2020), and conditioning on a baseline from the conditional probing of Hewitt et al. (2021), which isolates information beyond a baseline *representation*. Proposition 1 adds the accounting for the baseline *predictor*'s shortfall, the quantity that moves our numbers, and an audit that detects it. Yadavalli et al. (2026) report an order of magnitude more information in audio than in text, with context beyond the current sentence unavailable to the text side, the condition Proposition 1 describes. Appendix E places the rest of the lineage and sets both theorems against the closest prior bounds (Table 5).

5 Conclusion

We asked how much a spoken agent can gain by hearing and answered with a ceiling: I, the information the voice carries about the speaker's type beyond her words, caps the gain at $\sqrt{I/2}$ for one decision, and its per-turn terms cap the gain over a whole conversation. On television dialogue the words already carry 92 to 95 percent of what the voice says about the speaker's emotion, which puts the cap on an emotion-driven decision near 0.09. The with-and-without comparison credits audio with the text model's shortfall. Synthesis from an emotion label with instructions we wrote keeps a fifth to two fifths of what a person's delivery carries, and with a four-number channel on top, a simulator built on it cannot show what hearing is worth.

Three practices follow. A claim that a model hears should come with the ceiling and the share of it the model collects; on the same corpus one off-the-shelf model collects a third and another nothing detectable. What audio adds should be estimated against a words-only predictor as strong as the text model allows, and audited where a script fixes the words. Whether hearing helps should be tested on human speech, or on synthesis instructed as its makers document.

USE OF GENERATIVE AI

Large language models were used to revise prose, and to search the literature. Every claim of prior art in this paper was checked against the cited source by an author. Every formal result was checked by the Lean compiler, and the statements were reviewed separately against the text. The authors take responsibility for the content.

ETHICS STATEMENT

This work models a system that infers a user's emotional state from their voice in order to decide what to do next. Two risks follow. Inferred affect is a sensitive attribute, and a system that records it creates a record that did not exist before; the affect vector here enters the belief and is not retained beyond the conversation. And a system that reads affect in order to act on it can act against the user's interest as easily as for it, which is why success is scored against the user's own hidden goal rather than against a judge's impression, and why the deliberate stopping act exists. The results in Appendix A on chosen words show the other direction of the same concern: a user who learns what the voice does to the system may change how they speak, and the value of hearing can fall as a result.

We ran no human-subjects study. The simulated conversations use synthesized voices, but the benchmark measurements are made on recordings of human actors released under their corpora's own consent terms, and the paper reports results on those recordings. No proprietary or personal data enters the construction at any point.

REPRODUCIBILITY STATEMENT

Formal results. Every theorem, proposition and corollary of Section 2 and Appendix A is formalized in Lean 4 with Mathlib on finite types. The development builds with no `sorry` and no axioms beyond Mathlib's `propext`, `Classical.choice` and `Quot.sound`, printed from the build. Appendix B gives the formal statements, every difference from the text, and what is not formalized.

Numerical reproducibility. Aggregate results replay exactly from a fixed evaluation seed, and every ceiling reported agrees across two machines to machine precision; all fits are run to convergence. Episode-level planner artifacts reproduce on one machine, because decisions at near-indifference flip on floating-point variation (6 episodes in 300), so every episode-level comparison is run in one batch on one machine.

Compute and licenses. Every number was produced on serverless cloud infrastructure; the per-container billing logs total $92 over 149 containers on L40S and A100-80GB GPUs and on CPU. MELD is released under GPL-3.0 and its footage is copyrighted; RAVDESS is CC BY-NC-SA 4.0; CREMA-D is ODbL 1.0 with DbCL 1.0; Qwen2-Audio, parler-tts, CosyVoice3 and BERT-base are Apache-2.0; the bge sentence encoder is MIT; the audeering affect encoder is CC BY-NC-SA 4.0; WavLM-large and LLaMA-Omni 2 declare no license on their model repositories.

Release. The supplementary material holds the code, the Lean development and the result files behind every number, anonymized for review. Our trained weights and adapters will be released under an academic, non-commercial license; no base weights and no corpus audio are redistributed. The released affect head is fitted on all items, while every number reported for it is cross-fitted.

REFERENCES

András Antos, Luc Devroye, and László Györfi. Lower bounds for Bayes error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7):643–645, 1999. doi: 10.1109/34.777375.

Kenneth J. Arrow, David Blackwell, and M. A. Girshick. Bayes and minimax solutions of sequential decision problems. *Econometrica*, 17(3/4):213–244, 1949. doi: 10.2307/1905525.

Nicholas Asher and Alex Lascarides. Strategic conversation. *Semantics and Pragmatics*, 6(2):1–62, 2013. doi: 10.3765/sp.6.2.

Anton Bakhtin, David J. Wu, Adam Lerer, Jonathan Gray, Athul Paul Jacob, Gabriele Farina, Alexander H. Miller, and Noam Brown. Mastering the game of no-press Diplomacy via human-regularized reinforcement learning and planning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=F61FwJTZhb.

Ivan Balbuzanov. Lies and consequences: The effect of lie detection on communication outcomes. *International Journal of Game Theory*, 48(4):1203–1240, 2019. doi: 10.1007/s00182-019-00679-z.

Ken Binmore, Avner Shaked, and John Sutton. An outside option experiment. *The Quarterly Journal of Economics*, 104(4):753–770, 1989. doi: 10.2307/2937866.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Noam Brown, Tuomas Sandholm, and Brandon Amos. Depth-limited solving for imperfect-information games. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 7674–7685, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/34306d99c63613fad5b2a140398c0420-Abstract.html.

Trung H. Bui, Mannes Poel, Anton Nijholt, and Job Zwiers. A tractable hybrid DDN–POMDP approach to affective dialogue modeling for probabilistic frame-based dialogue systems. *Natural Language Engineering*, 15(2):273–307, 2009. doi: 10.1017/S1351324908005032.

Andrew Caplin and Mark Dean. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203, 2015. doi: 10.1257/aer.20140117.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113.

Yifu Chen, Shengpeng Ji, Zhengqing Liu, Qian Chen, Wen Wang, Ziqing Wang, Yangzhuo Li, Tianle Liang, and Zhou Zhao. Dual-axis generative reward model toward semantic and turn-taking robustness in interactive spoken dialogue models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 173–208. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.6.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, Hoboken, NJ, 2 edition, 2006.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982. doi: 10.2307/1913390.

Brendan Daley and Brett Green. Bargaining and news. *American Economic Review*, 110(2):428–474, 2020. doi: 10.1257/aer.20181316.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024. doi: 10.48550/ARXIV.2410.00037.

Wioletta Dziuda and Christian Salas. Communication with detectable deceit. SSRN working paper 3234695, 2018.

Paul Ekman and Wallace V. Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1): 88–106, 1969. doi: 10.1080/00332747.1969.11023575.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. LLaMA-Omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=PYmrUQmMEw.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. Llama-omni 2: LLM-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18617–18629, 2025b. doi: 10.18653/v1/2025.acl-long.912.

Leonardo Haw-Yang Foo, Chih-Kai Yang, Chen-An Li, Ke-Han Lu, and Hung-yi Lee. All that glitters is not audio: Rethinking text priors and audio reliance in audio-language evaluation. *arXiv preprint arXiv:2604.24401*, 2026. doi: 10.48550/ARXIV.2604.24401.

Alexander Frankel and Emir Kamenica. Quantifying information and uncertainty. *American Economic Review*, 109(10):3650–3680, 2019. doi: 10.1257/aer.20181897.

Drew Fudenberg and Jean Tirole. Sequential bargaining with incomplete information. *The Review of Economic Studies*, 50(2):221–247, 1983. doi: 10.2307/2297414.

Eitan Greenshtein. Comparison of sequential experiments. *Annals of Statistics*, 24(1):436–448, 1996. doi: 10.1214/aos/1033066220.

John C. Harsanyi. Games with incomplete information played by "Bayesian" players, i–iii part i. the basic model. *Management Science*, 14(3):159–182, 1967. doi: 10.1287/mnsc.14.3.159.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2333–2343. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1256.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. Conditional probing: Measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1626–1639, 2021. doi: 10.18653/v1/2021.emnlp-main.122.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi: 10.1109/TSSC.1966.300074.

Chi-Yuan Hsiao, Ke-Han Lu, Yu-Kuan Fu, Guan-Ting Lin, Hsiao-Tsung Hung, and Hung yi Lee. ASPIRin: Action space projection for interactivity-optimized reinforcement learning in full-duplex speech language models. *arXiv preprint arXiv:2604.10065*, 2026.

Dawei Huang, Yongjie Lv, Ruijie Xiong, Chunxiang Jin, and Xiaojiang Peng. When tone and words disagree: Towards robust speech emotion recognition under acoustic-semantic conflict. *arXiv preprint arXiv:2601.04564*, 2026.

Junichiro Ishida and Takashi Shimizu. Cheap talk with an informed receiver. *Economic Theory Bulletin*, 4(1):61–72, 2016. doi: 10.1007/s40505-015-0076-6.

Rebecca Jürgens, Kurt Hammerschmidt, and Julia Fischer. Authentic and play-acted vocal emotion expressions reveal acoustic differences. *Frontiers in Psychology*, 2:180, 2011. doi: 10.3389/fpsyg.2011.00180.

Patrik N. Juslin and Petri Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5):770–814, 2003. doi: 10.1037/0033-2909.129.5.770.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Navin Kartik. Strategic communication with lying costs. *Review of Economic Studies*, 76(4):1359–1395, 2009. doi: 10.1111/j.1467-937X.2009.00559.x.

Bhuvan Koduru, Dareen Safar B Alharthi, Rita Singh, and Bhiksha Raj. Heard but not heeded: Paralinguistic information encoding and loss in audio-language models. *arXiv preprint arXiv:2609.00727*, 2026.

Solomon Kullback. A lower bound for discrimination information in terms of variation. *IEEE Transactions on Information Theory*, 13(1):126–127, 1967. doi: 10.1109/TIT.1967.1053968.

Ernest K. Lai. Expert advice for amateurs. *Journal of Economic Behavior & Organization*, 103:1–16, 2014. doi: 10.1016/j.jebo.2014.03.023.

Petri Laukka, Daniel Neiberg, Mimmi Forsell, Inger Karlsson, and Kjell Elenius. Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. *Computer Speech & Language*, 25(1):84–104, 2011. doi: 10.1016/j.csl.2010.03.004.

Lucien Le Cam. Sufficiency and approximate sufficiency. *The Annals of Mathematical Statistics*, 35(4):1419–1455, 1964. doi: 10.1214/aoms/1177700372.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453, 2017. doi: 10.18653/v1/D17-1259.

Yuxin Li, Donghang Wu, Guan-Ting Lin, Hung yi Lee, Chengwei Qin, Zhehuai Chen, and Chen Chen. Decoupling conversational dynamics in full-duplex spoken models through reinforcement learning. *arXiv preprint arXiv:2607.07148*, 2026.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard J. Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying and modeling multimodal interactions: An information decomposition framework. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi: 10.1214/aoms/1177728069.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 3469–3483. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.269.

Xinyi Liu, Hooshang Nayyeri, Dilek Hakkani-Tur, Emine Yilmaz, Joo-Kyung Kim, Yifei Zhang, Charith Peris, and Hari Thadakamalla. Hear2Act: Benchmarking when prosody should change what an assistant does. *arXiv preprint arXiv:2608.19515*, 2026.

Yen-Ju Lu, Yuzhe Wang, Yaohan Guan, Xiluo He, Jiarui Hai, Mingrui Liang, Kaavya Chaparala, Thomas Thebaud, Laureano Moro-Velazquez, Najim Dehak, and Jesus Villalba. When text misleads: Inconsistent-aware reasoning for audio-grounded dialogue. *arXiv preprint arXiv:2608.27176*, 2026.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6150–6160, 2018.

Devon P. Major and Monita Chatterjee. Acoustic analyses of the RAVDESS corpus of emotional stimuli. *JASA Express Letters*, 6(2):024801, 2026. doi: 10.1121/10.0042364.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 875–884, 2020.

Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022. doi: 10.1126/science.ade9097.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017. doi: 10.1126/science.aam6960.

Inés Moreno de Barreda. Cheap talk with two-sided private information. *Games and Economic Behavior*, 148:97–118, 2024. doi: 10.1016/j.geb.2024.09.001.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087, 2020.

Roger B. Myerson and Mark A. Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 29(2):265–281, 1983. doi: 10.1016/0022-0531(83)90048-0.

Jiacheng Pang, Ashutosh Chaubey, and Mohammad Soleymani. Do audio LLMs listen or read? analyzing and mitigating paralinguistic failures with VoxParadox. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. doi: 10.48550/ARXIV.2605.27772.

Prashant Parikh. Communication and strategic inference. *Linguistics and Philosophy*, 14(5):473–514, 1991. doi: 10.1007/BF00632595.

Haim H. Permuter, Young-Han Kim, and Tsachy Weissman. Interpretations of directed information in portfolio theory, data compression, and hypothesis testing. *IEEE Transactions on Information Theory*, 57(6):3248–3259, 2011. doi: 10.1109/TIT.2011.2136270.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 527–536. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1050.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1961.

Tazeek Bin Abdur Rakib, Ambuj Mehrish, Lay-Ki Soon, Wern Han Lim, and Soujanya Poria. DialogXpert: Driving intelligent and emotion-aware conversations through online value-based reinforcement learning with LLM priors. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-26)*, pp. 29967–29975, 2026. doi: 10.1609/aaai.v40i36.40244.

Ariel Rubinstein. A bargaining model with incomplete information about time preferences. *Econometrica*, 53(5):1151–1172, 1985. doi: 10.2307/1911016.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016. URL http://jmlr.org/papers/v17/14-087.html.

Daniel Russo and James Zou. How much does your data exploration overfit? Controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2020. doi: 10.1109/TIT.2019.2945779.

Hitoshi Sadakane and Yin Chi Tam. Cheap talk and lie detection. *Games and Economic Behavior*, 158:787–814, 2026. doi: 10.1016/j.geb.2026.03.007.

Klaus R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1–2):227–256, 2003. doi: 10.1016/S0167-6393(02)00084-5.

Klaus R. Scherer. Vocal markers of emotion: Comparing induction and acting elicitation. *Computer Speech & Language*, 27(1):40–58, 2013. doi: 10.1016/j.csl.2011.11.003.

Daniel J. Seidmann. Effective cheap talk with conflicting interests. *Journal of Economic Theory*, 50 (2):445–458, 1990. doi: 10.1016/0022-0531(90)90012-9.

Christopher A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3): 665–690, 2003. doi: 10.1016/S0304-3932(03)00029-1.

Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.

Michael Spence. Job market signaling. *The Quarterly Journal of Economics*, 87(3):355–374, 1973. doi: 10.2307/1882010.

Siegfried Ludwig Sporer and Barbara Schwandt. Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology*, 20(4):421–446, 2006. doi: 10.1002/acp.1190.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. CIMA: A large open access dialogue dataset for tutoring. In Jill Burstein, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis, and Torsten Zesch (eds.), *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2020, Online, July 10, 2020*, pp. 52–64. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.BEA-1.5.

Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. PARADISE: A framework for evaluating spoken dialogue agents. In Philip R. Cohen and Wolfgang Wahlster (eds.), *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, 7-12 July 1997, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain*, pp. 271–280. Morgan Kaufmann Publishers / ACL, 1997. doi: 10.3115/976909.979652. URL https://aclanthology.org/P97-1035/.

Hui Wang, Fafa Zhang, Meng Liu, Xiangyu Chen, and Chaoxu Mu. UP-NRPA: user portrait based nested rollout policy adaptation for planning with large language models in goal-oriented dialogue systems. *arXiv preprint arXiv:2606.13683*, 2026a. doi: 10.48550/ARXIV.2606.13683.

Shuiyuan Wang, Zhixian Zhao, Hongfei Xue, Chengyou Wang, Shuai Wang, Hui Bu, Xin Xu, and Lei Xie. HumDial-EIBench: A human-recorded multi-turn emotional intelligence benchmark for audio language models. *arXiv preprint arXiv:2604.11594*, 2026b. doi: 10.48550/ARXIV.2604.11594.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63345–63354. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/wang25aw.html.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 5635–5649. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1566.

Ran Wei. Value of information and reward specification in active inference and POMDPs. *arXiv preprint arXiv:2408.06542*, 2024.

Shu wen Yang, Ming Tu, Andy T. Liu, Xinghua Qu, Hung yi Lee, Lu Lu, Yuxuan Wang, and Yonghui Wu. ParaS2S: Benchmarking and aligning spoken language models for paralinguistic-aware speech-to-speech interaction. In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*, 2026.

Jason D. Williams and Steve Young. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007. doi: 10.1016/j.csl.2006.06.008.

Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Wilcox, and Tamar Regev. Quantifying the redundancy between prosody and text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9765–9784, 2023. doi: 10.18653/v1/2023.emnlp-main.606.

Jiaqi Xiong, Yunjia Qi, Qi Cao, Yu Zheng, Yutong Zhang, Ziteng Wang, Ruofan Liao, Weisheng Xu, and Sichen Liu. DEAF: a benchmark for diagnostic evaluation of acoustic faithfulness in audio language models. *arXiv preprint arXiv:2603.18048*, 2026.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems 30*, pp. 2524–2533, 2017.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations (ICLR)*, 2020.

Aditya Yadavalli, Tiago Pimentel, Tamar I. Regev, Ethan Gotlieb Wilcox, and Alex Warstadt. What do prosody and text convey? Characterizing how meaningful information is distributed across multiple channels. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23665–23679, 2026. doi: 10.18653/v1/2026.acl-long.1085.

Denis Yarats and Mike Lewis. Hierarchical text generation and planning for strategic dialogue. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5591–5599. PMLR, 2018.

Steve Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013. doi: 10.1109/JPROC.2012.2225812.

Miron Zuckerman, Deborah T. Larrance, Nancy H. Spiegel, and Rafael Klorman. Controlling nonverbal displays: Facial expressions and tone of voice. *Journal of Experimental Social Psychology*, 17(5):506–524, 1981. doi: 10.1016/0022-1031(81)90037-8.

A PROOFS

Tools: for laws p, q on a finite set and payoffs u in $[0, 1]$, $|\mathbb{E}_p u - \mathbb{E}_q u| \leq \mathrm{TV}(p, q)$; Pinsker's inequality $\mathrm{TV}(p, q) \leq \sqrt{\frac{1}{2}\,\mathrm{KL}(p\|q)}$; and Jensen's inequality for the concave square root (Cover & Thomas, 2006).

A.1 THE BOUNDS OF SECTION 2

Theorem 1 (Value of hearing)**.** With $V(W) = \max_a \mathbb{E}[u(\theta, a) \mid W]$ and $V(W, Z) = \mathbb{E}[\max_a \mathbb{E}[u(\theta, a) \mid W, Z] \mid W]$, $0 \le \mathbb{E}[V(W, Z)] - \mathbb{E}[V(W)] \le \sqrt{\frac{1}{2} I(\theta; Z \mid W)}$; and if $\theta \perp\!\!\!\perp Z \mid (e, W)$ then $I(\theta; Z \mid W) \le I(e; Z \mid W)$, where e is the affect state.

Proof. Fix $W = w$; write $p_z = p(\theta \mid w, z)$ and $p = p(\theta \mid w)$. Lower bound: for the act a_w that is optimal under p, $\max_a \mathbb{E}_{p_z}[u(\cdot, a)] \ge \mathbb{E}_{p_z}[u(\cdot, a_w)]$, and $\mathbb{E}_{Z|w}\mathbb{E}_{p_Z}[u(\cdot, a_w)] = \mathbb{E}_p[u(\cdot, a_w)] = V(w)$. Upper bound: a difference of maxima is at most the maximum of the differences, so

$$V(w, Z) - V(w) \;\le\; \mathbb{E}_{Z|w}\Big[\max_a \big(\mathbb{E}_{p_Z}[u(\cdot, a)] - \mathbb{E}_p[u(\cdot, a)]\big)\Big] \;\le\; \mathbb{E}_{Z|w}\big[\mathrm{TV}(p_Z, p)\big]$$
$$\le\; \mathbb{E}_{Z|w}\sqrt{\tfrac{1}{2}\mathrm{KL}(p_Z \| p)} \;\le\; \sqrt{\tfrac{1}{2}\,\mathbb{E}_{Z|w}\mathrm{KL}(p_Z \| p)},$$

and $\mathbb{E}_{Z|w}\mathrm{KL}(p_Z \| p) = I(\theta; Z \mid W = w)$. Taking the expectation over W and applying Jensen once more gives the claim. For the second statement, $\theta \perp\!\!\!\perp Z \mid (e, W)$ gives $I(\theta, e; Z \mid W) = I(e; Z \mid W) + I(\theta; Z \mid e, W) = I(e; Z \mid W)$, and $I(\theta; Z \mid W) \le I(\theta, e; Z \mid W)$. □

Proposition 1 (What a fitted difference of log-losses measures, restated)**.** Let $\mathcal{L}(q) = \mathbb{E}[-\log q(\theta \mid \cdot)]$ be the expected log-loss of a conditional model of the type, with the conditioning set understood from the model. The best predictor from the words and the best from the words and the voice are the true conditionals, so $L^*_W - L^*_{WZ} = I(\theta; Z \mid W)$. For fitted predictors with losses L_W and L_{WZ}, shortfalls $\mathrm{ex}' = L_W - L^*_W \ge 0$ and $\mathrm{ex} = L_{WZ} - L^*_{WZ} \ge 0$, and difference $D = L_W - L_{WZ}$,

$$D \;=\; I(\theta; Z \mid W) \;+\; \mathrm{ex}' \;-\; \mathrm{ex}.$$

Proof. For a conditional model q of θ given a conditioning variable C, $\mathcal{L}(q) - \mathcal{L}(p_{\theta|C}) = \mathbb{E}\big[\mathrm{KL}(p_{\theta|C} \,\|\, q(\cdot \mid C))\big] \ge 0$, with equality if and only if $q = p_{\theta|C}$ almost surely; this is Gibbs' inequality applied conditionally. So each infimum over all conditional laws is attained at the true conditional, and $\inf_{q'} \mathcal{L}(q') - \inf_q \mathcal{L}(q) = \mathbb{E}[-\log p(\theta \mid W)] - \mathbb{E}[-\log p(\theta \mid W, Z)] = \mathbb{E}\big[\log \frac{p(\theta|W,Z)}{p(\theta|W)}\big] = I(\theta; Z \mid W)$. For the restricted classes, $D = I(\theta; Z \mid W) + \mathrm{ex}(\mathcal{Q}') - \mathrm{ex}(\mathcal{Q})$, and both excess losses are nonnegative. A weak words-only class inflates the estimate by its own excess loss, so the words-only predictor must be at least as strong as the words pathway of the words-and-voice predictor. □

Proof of Corollary 2. The best words-only predictor attains the conditional entropy, so $\mathrm{ex}' = L_W - \mathrm{H}(\theta \mid W)$ and $D - \mathrm{ex}' = \mathrm{H}(\theta \mid W) - L_{WZ}$. No predictor from (W, Z) has expected log-loss below $\mathrm{H}(\theta \mid W, Z)$, by Gibbs' inequality, so $D - \mathrm{ex}' \le \mathrm{H}(\theta \mid W) - \mathrm{H}(\theta \mid W, Z) = I(\theta; Z \mid W)$. If the words carry nothing about the type, $\mathrm{H}(\theta \mid W) = \mathrm{H}(\theta)$. A words-only predictor that ignores W predicts one law q on every item, and its loss on the test labels is the cross-entropy of their empirical law $\hat{p}$ against q, which is at least $\mathrm{H}(\hat{p})$; with the label entropy computed on the same labels, $\mathrm{ex}' \ge 0$ holds exactly, and $\mathrm{ex}' < 0$ shows that the predictor's output varies with W. □

Proof of Theorem 2. Let P be the law of the whole conversation under π, hidden states included, and let Q be its law when π is fed, at each turn t, a fake voice drawn from $\bar{p}(\cdot \mid \tilde{S}_t)$, the law of z_t given S_t under π evaluated at the history $\tilde{S}_t$ that π has been fed, on the histories π reaches and any fixed law elsewhere. Under Q, π plays a deaf policy, so $\mathbb{E}_Q[G] \le V^*_W$, while $\mathbb{E}_P[G] = V(\pi)$. The two laws share every kernel given the past, for the hidden state, the words and the acts, and differ only in the law of each voice: under P it is $p_z(\cdot \mid x_{t+1}, h_t, a_t, w_t)$, which depends on the past only through (x_{t+1}, S_t), and under Q it is $\bar{p}(\cdot \mid S_t)$. The chain rule for the Kullback–Leibler divergence therefore gives

$$\mathrm{KL}(P \,\|\, Q) \;=\; \sum_{t=1}^{H-1} \mathbb{E}_P\Big[\mathbf{1}[\tau > t]\,\mathrm{KL}\big(p_z(\cdot \mid x_{t+1}, S_t) \,\big\|\, \bar{p}(\cdot \mid S_t)\big)\Big].$$

Given $S_t = s$, the law $\bar{p}(\cdot \mid s)$ is the mixture of $p_z(\cdot \mid x, s)$ over the posterior of x_{t+1}, so the mean of the inner divergence over that posterior is the mutual information between x_{t+1} and z_t

given $S_t = s$; hence $\mathrm{KL}(P \,\|\, Q) = \sum_t \iota_t$. The return lies in an interval of width R, so $\mathbb{E}_P[G] - \mathbb{E}_Q[G] \le R\,\mathrm{TV}(P, Q) \le R\sqrt{\tfrac{1}{2}\mathrm{KL}(P \,\|\, Q)}$ by Pinsker's inequality, which is the claim. This is the first inequality. The second holds because the square root is subadditive, and the Cauchy–Schwarz inequality is not needed. Every deaf policy is a hearing policy, which gives $V^*_{WZ} \ge V^*_W$. □

A.2 Sharpness, continuous voices, and conversations

Proposition A.1 (The constant is sharp)**.** *Let θ be uniform on $\{0, 1\}$, let the words carry nothing, let $Z \in \{0, 1\}$ equal θ with probability $\frac{1}{2} + \delta$ for some $\delta \in (0, \frac{1}{2}]$, and let two acts pay $u(\theta, a) = \mathbf{1}[a = \theta]$. Then $\mathbb{E}[V(W, Z)] - \mathbb{E}[V(W)] = \delta$ and $2\delta^2 \le I(\theta; Z \mid W) \le 4\delta^2$, so*

$$\tfrac{1}{\sqrt{2}}\sqrt{\tfrac{1}{2}I(\theta; Z \mid W)} \;\le\; \mathbb{E}[V(W, Z)] - \mathbb{E}[V(W)] \;\le\; \sqrt{\tfrac{1}{2}I(\theta; Z \mid W)}.$$

Moreover the ratio of the gain to the bound tends to one as $\delta \to 0$.

Proof of Proposition A.1. Under the prior both acts pay $\frac{1}{2}$, so $V(W) = \frac{1}{2}$. Given $Z = z$ the posterior puts $\frac{1}{2} + \delta$ on $\theta = z$, so $V(W, Z) = \frac{1}{2} + \delta$ and the gain is δ. Since Z is uniform, $I(\theta; Z) = \mathrm{KL}(\mathrm{Bern}(\frac{1}{2} + \delta) \,\|\, \mathrm{Bern}(\frac{1}{2}))$. Pinsker's inequality gives the lower bound $2\,\mathrm{TV}^2 = 2\delta^2$. The upper bound follows from $\log y \le y - 1$ applied to each term of the divergence: $(\frac{1}{2} + \delta)\log(1 + 2\delta) + (\frac{1}{2} - \delta)\log(1 - 2\delta) \le (\frac{1}{2} + \delta)2\delta - (\frac{1}{2} - \delta)2\delta = 4\delta^2$. The ratio of the gain to the bound is $\delta/\sqrt{I/2}$, and it tends to one because $I/(2\delta^2) \to 1$ as $\delta \to 0$: the divergence is a smooth function of δ whose value and first derivative vanish at 0 and whose second derivative there is 4. □

Proposition A.2 (The constant of Theorem 2 is sharp)**.** *For every $\eta > 0$ there is a two-channel conversation problem with $H = 2$ in which $V^*_{WZ} - V^*_W \ge (1 - \eta)\, R\sqrt{\iota_1/2}$, with ι_1 taken under the optimal hearing policy.*

Proof. Make the instance of Proposition A.1 one listening turn. Two types are equally likely, there is one affect state, and there is one word, so the words carry nothing. The one continuing act, ask, is answered by a voice $z_1 \in \{0, 1\}$ that equals the type with probability $\frac{1}{2} + \delta$. The ending acts are a_0 and a_1, with $r(\theta, a_i) = \mathbf{1}[i = \theta]$, and fold, with $\rho \equiv 0$. Let $H = 2$ and $c = \delta^2$, so $R = 1 + \delta^2$. A deaf policy learns nothing by asking, so its best is to end at once with a_0 or a_1, and $V^*_W = \frac{1}{2} - c$. A hearing policy that asks and then takes the act the voice names earns $\frac{1}{2} + \delta - 2c$, which exceeds every alternative because $\delta > c$ for $\delta < 1$; so it is optimal and $V^*_{WZ} - V^*_W = \delta - \delta^2$. Under it the conversation always survives the first turn, S_1 is fixed, and the hidden state after the turn is the type, so $\iota_1 = I(\theta; z_1) = \mathrm{KL}(\mathrm{Bern}(\frac{1}{2} + \delta) \,\|\, \mathrm{Bern}(\frac{1}{2}))$. By the limit in the proof of Proposition A.1, $\iota_1/(2\delta^2) \to 1$, so $R\sqrt{\iota_1/2}/\delta \to 1$ and the ratio $(\delta - \delta^2)/(R\sqrt{\iota_1/2})$ tends to one as $\delta \to 0$. □

Corollary A.3 (Act-information bound)**.** *In the setting of Theorem 1, let A^* be an act the hearing system takes, that is, a maximizer of $\mathbb{E}[u(\theta, a) \mid W, Z]$, so that $V(W, Z) = \mathbb{E}[u(\theta, A^*) \mid W, Z]$. Then*

$$\mathbb{E}[V(W, Z)] - \mathbb{E}[V(W)] \;\le\; \sqrt{\tfrac{1}{2}\, I(\theta; A^* \mid W)} \;\le\; \sqrt{\tfrac{1}{2}\, I(\theta; Z \mid W)},$$

and the middle term is at most $\sqrt{\frac{1}{2}\log|A|}$ for an act set A. The same argument applied to an arbitrary function of (W, Z) bounds the gain of that act alone, not the gain of the best one; a constant act carries no information about the type while the best act may gain a great deal.

Proof of Corollary A.3. Given $W = w$, let θ' be drawn from $p(\theta \mid w)$ independently of A^*. Then $\mathbb{E}[V(W, Z)] = \mathbb{E}[u(\theta, A^*)]$, and $\mathbb{E}[V(W)] \ge \mathbb{E}[u(\theta', A^*)]$, since each act A^* names is available to the system that has only the words. The difference $\mathbb{E}[u(\theta, A^*)] - \mathbb{E}[u(\theta', A^*)]$ compares, given w, the joint law of (θ, A^*) with the product of its marginals, so the total-variation bound, Pinsker's inequality, and Jensen's inequality, as in the proof of Theorem 1, bound it by $\sqrt{\frac{1}{2}I(\theta; A^* \mid W)}$. Data processing gives the second inequality, since A^* is a function of (W, Z), and $I(\theta; A^* \mid W) \le \mathrm{H}(A^* \mid W) \le \log|A|$ gives the last. □

Proposition A.4 (Hearing through an encoder)**.** *In the setting of Theorem 1, let a system hear the voice only through an encoder f, a function of the words and the voice with values in a finite set. It gains at most $\sqrt{\frac{1}{2}\, I(\theta; f(W,Z) \mid W)}$, and $I(\theta; f(W,Z) \mid W) \le I(\theta; Z \mid W)$.*

Proof. The first claim is Theorem 1 with $f(W,Z)$ in place of Z. Given $W = w$, $f(w,Z)$ is a function of Z, so the data-processing inequality for the Kullback–Leibler divergence (Cover & Thomas, 2006), applied to the map $(\theta, z) \mapsto (\theta, f(w,z))$, gives $I(\theta; f(W,Z) \mid W = w) \le I(\theta; Z \mid W = w)$ for every w, and averaging over w gives the second claim. Corollary A.3 is the case in which f is the act the hearing system takes. □

Corollary A.5 (From utterances to conversations)**.** *In a two-channel conversation problem, suppose the voice depends on the past only through the current hidden state, the act and the words: at turn t, $p_z(\cdot \mid x', h, a, w) = q_t(\cdot \mid x', a, w)$. Fix any hearing policy π, let $m_t(a,w) = P_\pi(\tau > t, a_t = a, w_t = w)$ be the probability that the conversation reaches turn t, continues, and shows act a and words w, and let $J_t = I(x_{t+1}; z_t \mid a_t, w_t)$ under the law of the turns π reaches, $P_\pi(\cdot \mid \tau > t)$. Then for every $t < H$*

$$\iota_t \;\le\; \sum_{a,w} m_t(a,w)\, I(x_{t+1}; z_t \mid a_t = a, w_t = w) \;=\; P_\pi(\tau > t)\, J_t,$$

and so

$$V(\pi) - V^*_W \;\le\; R \sum_{t=1}^{H-1} \sqrt{\tfrac{1}{2}\, P_\pi(\tau > t)\, J_t}.$$

If moreover the type reaches the voice only through the affect, $q_t(\cdot \mid (\theta, e'), a, w) = q_t(\cdot \mid e', a, w)$, then $J_t = I(e_{t+1}; z_t \mid a_t, w_t)$ under the same law.

Proof. The event $\{\tau > t\}$ says that $a_1, \dots, a_t$ are continuing acts, so it is determined by S_t, and conditioning on it leaves the law of (x_{t+1}, z_t) given S_t unchanged. Hence $\iota_t = P_\pi(\tau > t)\, I_Q(x_{t+1}; z_t \mid S_t)$ with $Q = P_\pi(\cdot \mid \tau > t)$. Write $S_t = (U_t, a_t, w_t)$ with $U_t = (h_t, z_{1:t-1})$. By Definition 1 and the assumption, the law of z_t given (x_{t+1}, U_t, a_t, w_t) is $q_t(\cdot \mid x_{t+1}, a_t, w_t)$, under P_π and under Q alike, so z_t is independent of U_t given (x_{t+1}, a_t, w_t). Two expansions of the chain rule give

$$\begin{aligned} I_Q(x_{t+1}; z_t \mid U_t, a_t, w_t) \;&\le\; I_Q(x_{t+1}, U_t; z_t \mid a_t, w_t) \\ &=\; I_Q(x_{t+1}; z_t \mid a_t, w_t) + I_Q(U_t; z_t \mid x_{t+1}, a_t, w_t) \;=\; J_t, \end{aligned}$$

the last term being zero, and substituting into Theorem 2 gives the display. Under the second assumption, z_t is independent of θ given (e_{t+1}, a_t, w_t) under Q, so $J_t = I_Q(e_{t+1}; z_t \mid a_t, w_t) + I_Q(\theta; z_t \mid e_{t+1}, a_t, w_t) = I_Q(e_{t+1}; z_t \mid a_t, w_t)$. □

Corollary A.6 (The ceiling moves continuously with the population)**.** *Let a second population share the prior μ, the acts and the payoffs of a two-channel conversation problem but follow other user laws T', p'_w, p'_z, and suppose that at every turn and every (x_t, h_t, a_t) the two laws of the user's response (x_{t+1}, w_t, z_t) differ by at most ε in total variation. Write V' for values under the second population and ι_t for the terms of Theorem 2 under the first. Then for every hearing policy π,*

$$V'(\pi) - V'^*_W \;\le\; R \sum_{t=1}^{H-1} \sqrt{\tfrac{1}{2}\, \iota_t} \;+\; 2R(H-1)\,\varepsilon.$$

Proof. Fix any policy σ, hearing or deaf, and let P and P' be the laws of its conversations under the two populations. For $k = 0, \dots, H-1$ let $P^{(k)}$ use the second population's response law at turns $t \le k$ and the first's after, so $P^{(0)} = P$ and $P^{(H-1)} = P'$. Consecutive laws share every kernel except the response law at turn k, so $\mathrm{TV}(P^{(k-1)}, P^{(k)}) \le \varepsilon$, since total variation does not grow when a common kernel is applied. Hence $\mathrm{TV}(P, P') \le (H-1)\varepsilon$, and because the return lies in an interval of width R, $|V'(\sigma) - V(\sigma)| \le R(H-1)\varepsilon$. Applying this to π and to an optimal deaf policy σ_W of the first population gives $V'(\pi) - V'^*_W \le V'(\pi) - V'(\sigma_W) \le V(\pi) - V^*_W + 2R(H-1)\varepsilon$, and Theorem 2 bounds $V(\pi) - V^*_W$. □

Corollary A.7 (When to stop listening). *Let $\bar{\iota}$ be the most that one turn can say about the hidden state: the supremum, over pre-turn histories $s = (h_t, z_{1:t-1})$ and continuing acts a, of $I(x_{t+1}; (W_t, Z_t) \mid (h_t, z_{1:t-1}) = s, a_t = a)$. Suppose further that the ending payoffs do not drift under the dynamics, which holds here because the type does not move and an ending payoff reads the type. If $\sqrt{\bar{\iota}/2} \le c$, then at every history an ending act is optimal, for hearing and deaf policies alike.*

Proof of Corollary A.7. Let $W_n(s)$ be the best expected return from history s when at most n more continuing acts may be taken, so that $W_0(s) = \max_{a \in A_e} \mathbb{E}[r(\theta, a) \mid s] - ct$ at turn t. Ending payoffs depend on the hidden state through θ alone, so Theorem 1, with the ending acts as the acts and the turn's words and voice as the observation, gives $\mathbb{E}[W_0(S_{t+1}) \mid s, a] \le W_0(s) + \sqrt{\bar{\iota}/2} - c$ for every continuing act a. Since $W_n(s) = \max\{W_0(s), \max_a \mathbb{E}[W_{n-1}(S_{t+1}) \mid s, a]\}$, induction on n gives $W_n(s) \le W_0(s) + n\,(\sqrt{\bar{\iota}/2} - c)^+$. When $\sqrt{\bar{\iota}/2} \le c$ the right side is $W_0(s)$. □

Proposition A.8 (Shape of the fold region). *Fix a turn $t < H$ and a spoken history h, and let b range over beliefs on X, the state at turn t being drawn from b. Let $F_t(h)$ be the set of beliefs at which fold is an optimal act for a hearing policy. (i) $F_t(h)$ is convex. (ii) If every completing act fails on every out-of-reach type, $F_t(h)$ contains the point mass on each out-of-reach state; if each in-reach type has a completing act that succeeds on it, $F_t(h)$ contains no point mass on an in-reach state. (iii) If the voice Z' is a garbling of Z, that is, z' is drawn from a kernel $\kappa(\cdot \mid z)$ that does not depend on the state, then $F_t(h; Z) \subseteq F_t(h; Z')$: a better voice folds at fewer beliefs. (iv) If the user's laws depend on the history only through a summary that does not count turns, the region grows as the end approaches: with n turns remaining and $F(n)$ the region at a fixed summary, $F(n) \subseteq F(n-1)$.*

Proof of Proposition A.8. (i) A continuation strategy σ for turns $t, \dots, H$ maps what will be seen to acts and does not depend on b, so its expected return is linear in b: $\langle b, \alpha_\sigma \rangle$ with $\alpha_\sigma(x)$ the expected return from state x. There are finitely many strategies. The best return from a continuing act, $V_t^{\mathrm{c}}(h, b) = \max_\sigma \langle b, \alpha_\sigma \rangle$ over strategies that continue at turn t, is therefore a maximum of linear functions of b and convex; the return of an ending act a is $\langle b, r(\cdot, a) \rangle - ct$, linear. Fold is optimal exactly where $\langle b, \rho \rangle - ct \ge V_t^{\mathrm{c}}(h, b)$ and $\langle b, \rho \rangle \ge \langle b, r(\cdot, a) \rangle$ for every other ending act a. The first set is where a convex function lies below a linear one, hence convex; the others are half-spaces; the intersection is convex. (ii) At the point mass on an out-of-reach state, fold returns $1 - ct$, every other ending act returns $-ct$, and every strategy that continues ends at some $\tau > t$ with return at most $1 - c\tau < 1 - ct$; so fold is uniquely optimal. At the point mass on an in-reach state with a completing act that succeeds on it, that act returns $1 - ct > \rho_{\mathrm{in}} - ct$, so fold is not optimal. (iii) A strategy for Z' becomes a strategy for Z by drawing z' from $\kappa(\cdot \mid z)$ itself, with the same expected return, because the kernel does not depend on the state. So $V_t^{\mathrm{c}}(h, b; Z) \ge V_t^{\mathrm{c}}(h, b; Z')$, and the fold condition under Z implies it under Z'. (iv) Net of the cost already paid, a strategy for $n - 1$ remaining turns is a strategy for n that never uses the last turn and earns the same return, since the laws do not see the turn count; so $V^{\mathrm{c}}(n) \ge V^{\mathrm{c}}(n-1)$ at the same summary and belief, and the fold condition with n turns left implies it with $n - 1$. □

A.3 When the words are chosen

The definition fixes the law of the words. This subsection lets the user choose them against a system that hears, which is where the fixed-population results stop and the game view is needed. The results cut both ways.

Proposition A.9 (Chosen words can undo hearing). *Take one turn with three types under the prior $(0.4, 0.35, 0.25)$, three acts a_1, a_2, a_3, and a binary voice with $P(z = \mathrm{H} \mid \theta_3) = q$ and $P(z = \mathrm{H} \mid \theta_1) = P(z = \mathrm{H} \mid \theta_2) = 1 - q$ for some $q \in [\frac{1}{2}, 1]$. The system is paid 1 when a_i meets θ_i and 0 otherwise. The types are paid $(1, \frac{1}{2}, 0)$, $(\frac{9}{10}, 1, 0)$, and $(0, 1, \frac{1}{2})$ for (a_1, a_2, a_3): type 3 wants what type 2 wants. The user chooses one of two messages m_1, m_2; the system sees the message and the voice. (a) For $q \le \frac{7}{12}$, the profile in which θ_1 sends m_1, θ_2 and θ_3 send m_2, and the system answers a_1 and a_2 whatever the voice, is a perfect Bayesian equilibrium and pays the system 0.75. (b) For $\frac{7}{12} < q < \frac{9}{10}$ no equilibrium has that message structure, while the profile in which θ_1 and θ_2 send m_1, θ_3 sends m_2, and the system answers a_1 and a_3 is an equilibrium; it pays the system 0.65 and uses no voice. (c) For $q \ge \frac{9}{10}$ the profile of (a) with the system answering a_3 on (m_2, H) is an equilibrium and pays the system $0.4 + 0.6q$.*

Proof of Proposition A.9. Posteriors: m_1 in (a) and (c) comes from θ_1 alone, so a_1 is the system's best reply to it. On (m_2, L) in (a) and (c) the posterior weights are $0.35q$ on θ_2 and $0.25(1-q)$ on θ_3, so a_2 is best for every $q \geq \frac{1}{2}$. On (m_2, H) they are $0.35(1-q)$ and $0.25q$: a_2 is best when $q \leq \frac{7}{12}$ and a_3 is strictly best when $q > \frac{7}{12}$. In (b), m_1 carries weights 0.4 and 0.35, so a_1 is best, and the voice, which has the same law under θ_1 and θ_2, changes nothing; m_2 comes from θ_3 alone, so a_3 is best. Deviations: in (a) each type receives its favorite act, so none deviates. In (c), θ_1 receives 1; θ_2 receives a_2 when $z = \mathrm{L}$, which has probability q, so $q \geq \frac{9}{10}$ against $\frac{9}{10}$ from m_1; θ_3 receives $(1-q) + \frac{q}{2}$ against 0 from m_1. For the first claim of (b), any equilibrium with that message structure has the system answer a_3 on (m_2, H), since that reply is strictly best, and a_1 on m_1; then θ_2 receives $q < \frac{9}{10}$ from m_2 and $\frac{9}{10}$ from m_1, and deviates. For the second, θ_1 receives 1; θ_2 receives $\frac{9}{10}$ against 0 from m_2; θ_3 receives $\frac{1}{2}$ against 0 from m_1. The system's payoffs are the prior weights of the types it answers correctly: $0.4 + 0.35$, $0.4 + 0.25$, and $0.4 + 0.35q + 0.25q$. □

Theorem A.1 (Two kinds of voice)**.** *Take one turn with finite types Θ under a full-support prior, acts A, system payoff $u(\theta, a)$, user payoff $v(\theta, a)$, and messages that are nonempty subsets of Θ, a message w being true for θ when $\theta \in w$. A user strategy is truthful when every message a type sends is true for it. A voice* hears the type *when $z \sim p(z \mid \theta)$ whatever the message, and* hears the lie *at rate p when $z \in \{\mathrm{clean}, \mathrm{flag}\}$ and $P(\mathrm{flag} \mid \theta, w) = p\,\mathbf{1}[\theta \notin w]$. An equilibrium is a perfect Bayesian equilibrium in which, at a flagged message that no type sends, the system's reply is a best reply to some belief on the types outside the message, since a flag proves the message false. A truthful equilibrium (σ, α) of the words-only game* admits punishments *if for every message $w \subsetneq \Theta$ there is an act a^-_w, a best reply to some belief on $\Theta \setminus w$, with $v(\theta, a^-_w) \leq v(\theta, \alpha(w))$ for every $\theta \notin w$. The message Θ needs no punishment: it is true for every type, so no type can lie by sending it and the flag never fires on it.*

(a) A voice that hears the type can remove a truthful equilibrium, and the system's payoff in the best truthful equilibrium need not be monotone in the voice's accuracy (Proposition A.9).

(b) Let (σ, α) be a truthful equilibrium of the words-only game that admits punishments. For every $p \in [0, 1]$, the profile that keeps σ, answers $\alpha(w)$ to a clean w and a^-_w to a flagged w is an equilibrium of the game with a voice that hears the lie at rate p, with the same outcome and the same payoffs. A truthful profile of this form that is an equilibrium at rate p is one at every rate $p' \geq p$; so the set of such equilibria grows with p, and the system's best payoff over them is nondecreasing in p.

(c) Let messages be type claims, let $a^*(\theta)$ be a best reply to the point mass on θ, and let a^-_θ be a best reply to some belief on $\Theta \setminus \{\theta\}$ with $v(\theta', a^-_\theta) \leq v(\theta', a^*(\theta'))$ for every $\theta' \neq \theta$. Full revelation, each type claiming itself, answered by a^* when clean and by a^- when flagged, is an equilibrium iff $p \geq p^*$, where

$$p^* \;=\; \max_{\theta \neq \theta' \colon v(\theta, a^*(\theta')) > v(\theta, a^*(\theta))} \frac{v(\theta, a^*(\theta')) - v(\theta, a^*(\theta))}{v(\theta, a^*(\theta')) - v(\theta, a^-_{\theta'})} \;\in\; [0, 1],$$

and $p^* = 0$ when no type prefers another type's reply to its own.

Proof of Theorem A.1. (a) is Proposition A.9: the voice there has the same law under every message; the profiles are truthful once m_1 and m_2 are read as the sets of types that send them; the equilibrium worth 0.75 exists at $q \leq \frac{7}{12}$ and not at $q \in (\frac{7}{12}, \frac{9}{10})$; and the payoffs 0.75, 0.65, and $0.4 + 0.6q$ are not monotone in q.

(b) Under a truthful σ no type is flagged, so a clean message carries exactly what it carried in the words-only game: the posterior at (w, clean) is the words-only posterior, and $\alpha(w)$ remains a best reply. A flagged w has probability zero, and a^-_w is a best reply to a belief on $\Theta \setminus w$ by assumption. A type θ that sends a true message is never flagged and receives $v(\theta, \alpha(w))$, as before. If it sends a false w' it receives $(1-p)\,v(\theta, \alpha(w')) + p\,v(\theta, a^-_{w'}) \leq v(\theta, \alpha(w'))$, which the words-only equilibrium already bounds by the payoff of a true message. So every constraint holds, no flag sounds, and all payoffs are unchanged. The system's constraints do not involve p, and the user's constraint against a false w' has a left side nonincreasing in p, because $v(\theta, a^-_{w'}) \leq v(\theta, \alpha(w'))$; a profile that satisfies it at p satisfies it at every larger rate. A maximum over a growing family is nondecreasing.

(c) At a clean claim the posterior is the point mass, so a^* is a best reply; a flagged claim has probability zero, and a^- is a best reply to a belief outside the claim. Type θ claiming itself receives $v(\theta, a^*(\theta))$; claiming θ' it receives $(1-p)\,v(\theta, a^*(\theta')) + p\,v(\theta, a^-_{\theta'})$. The false claim is unprofitable

iff $p\,[v(\theta, a^*(\theta')) - v(\theta, a^-_{\theta'})] \geq v(\theta, a^*(\theta')) - v(\theta, a^*(\theta))$. When the right side is at most zero this holds for every p, since the bracket is nonnegative. Otherwise the bracket is at least the right side, by the assumption on a^-, so the ratio lies in $(0, 1]$ and the condition is that p be at least the ratio. The maximum over pairs gives p^*. □

Proposition A.10 (Listening to a voice that hears the type)**.** *Let $m \geq 2$ types have prior μ, let the system observe n voices drawn independently from $p(\cdot \mid \theta)$, and let $\rho \geq \sum_z \sqrt{p(z \mid \theta)\, p(z \mid \theta')}$ for every pair of distinct types. The most probable type given $z_{1:n}$ is wrong with probability at most $\frac{m-1}{2}\,\rho^{\,n}$, and for payoffs $r(\theta, a) \in [0, 1]$ the best act given $z_{1:n}$ earns at least $\mathbb{E}[\max_a r(\theta, a)] - \frac{m-1}{2}\,\rho^{\,n}$.*

Proof. Let $\hat{\theta}$ be a most probable type given $z_{1:n}$ and $a^*(\theta)$ an act maximizing $r(\theta, \cdot)$. Then $\mathbb{E}[\max_a \mathbb{E}[r(\theta, a) \mid z_{1:n}]] \geq \mathbb{E}[r(\theta, a^*(\hat{\theta}))] \geq \mathbb{E}[\max_a r(\theta, a)] - P(\hat{\theta} \neq \theta)$, because payoffs lie in $[0, 1]$ and the two acts agree when $\hat{\theta} = \theta$. The error of the most probable type is $\sum_{z_{1:n}} [\sum_\theta \mu(\theta) p(z_{1:n} \mid \theta) - \max_\theta \mu(\theta) p(z_{1:n} \mid \theta)]$, and for nonnegative numbers the sum minus the maximum is at most the sum of the pairwise minima, so it is at most $\sum_{\theta<\theta'} \sum_{z_{1:n}} \min\{\mu(\theta) p(z_{1:n} \mid \theta), \mu(\theta') p(z_{1:n} \mid \theta')\}$. Each minimum is at most the geometric mean, and the geometric means factor over the n independent voices, which gives at most $\sum_{\theta<\theta'} \sqrt{\mu(\theta)\mu(\theta')}\,\rho^{\,n} \leq \frac{m-1}{2}\rho^{\,n}$, the last step because $\sum_{\theta<\theta'} \sqrt{\mu(\theta)\mu(\theta')} = \frac{1}{2}[(\sum_\theta \sqrt{\mu(\theta)})^2 - 1] \leq \frac{m-1}{2}$ by the Cauchy–Schwarz inequality. □

The two signs are the point. A voice that reads the type gives the user a reason to stop distinguishing herself by what she says, and the system's payoff is not monotone in how good that voice is. A voice that catches lies never costs the system an equilibrium it already had, and buys full revelation once it is accurate enough. In the problem of Definition 1 the natural punishment is already present: fold is a best reply when the flagged words leave only out-of-reach types, and every in-reach user prefers an honest answer to being folded on. The voice is therefore better used to check the words than to guess the type.

Relation to cheap talk with lie detection. The three phenomena above are known in a more general message space. Balbuzanov (2019) studies a sender-receiver game with an exogenous detection probability p and shows that fully revealing equilibria exist well below $p = 1/2$, that with precise messages they exist only at $p = 1$ and on an intermediate set bounded away from both ends, so that the receiver's best equilibrium payoff is *non-monotone* in p, and that admitting vague messages restores full revelation for all large enough p. Part (a) here is that non-monotonicity, for a voice that reads the type rather than one that checks a claim. Parts (b) and (c) are monotone, and the reason is the restriction: our messages are type claims and our equilibria admit punishments, which is the case in which Balbuzanov's non-monotonicity does not arise. We do not claim monotonicity in general, and a reader should take (b) and (c) as statements about that restricted class rather than about lie detection at large. What the restriction buys is that it is the class a spoken system actually faces, since fold is a punishment that is always available and every in-reach user prefers an honest answer to being folded on.

B The Machine-Checked Statements

The results of Section 2 are formalized in Lean 4 with Mathlib on finite types. The development builds with no `sorry`; a script prints the axioms of every theorem named below from the build, and each depends on `propext`, `Classical.choice` and `Quot.sound` and nothing else. This appendix gives the formal statements of the four results the measurements rest on, the map from the paper's results to the Lean theorems, and every place the formal statement differs from the text.

Map. Theorem 1 is `value_of_hearing`; its second clause is `CMI_le_of_factor`. Theorem 2 is `OneShot.oneshot_bound` for its first inequality, with the chain-rule identity `OneShot.klTop_eq_iotaTop`, and `conversation_boundR` for the per-turn form. Proposition 1 is `estimate_identity` with `info_ge_corrected`; Corollary 2

is `corrected_exact` with `vacuous_words_measurable`. The sharpness of the constant is `Sharp.sharp` and `Sharp.sharp_ratio_tendsto_one`. Proposition A.2 is `SharpConv.sharpconv`, with the limit `Sharp.conv_ratio_tendsto_one`, stated for the hearing policy of its proof, whose optimality is not needed because Theorem 2 holds for every policy. Corollary A.6 is `conversation_bound_drift`, with the total-variation hypothesis on the joint response (x_{t+1}, w_t, z_t) given the state, the spoken history and the act. Corollary A.5 is `UtteranceProcess.iotaTopR_le_mass_mul_mi` with the affect clause `UtteranceProcess.iotaTopR_le_affect`, resting on the finite-law core `Utterance.info_le_of_local` and `Utterance.info_eq_affect`; the chain-rule bound of Section 2 on the affect and the type is `AffectLeak.type_le_affect_add`. For Proposition A.4 with a finite voice, `value_of_hearing`, `valueWZ_sub_valueW_le_CMIA` and `CMIA_le_CMI` give the act-quantization bound and the encoder bound. Proposition A.10 is `Listening.Pe_le_card` and `Listening.valueListen_ge`. Pinsker's inequality, which Mathlib lacks, is proved in the development.

Statements. Printed in ASCII; the source uses Unicode. `IsDist` asserts a probability vector, `CMI` the conditional mutual information $I(\theta; Z \mid W)$ on finite types, and `valueW`, `valueWZ` the two values of Theorem 1.

```
theorem value_of_hearing [Fintype W] {pw : W -> Real} {pz : W -> Z -> Real}
    {pt : W -> Z -> Theta -> Real} {u : Theta -> A -> Real}
    (hpw : IsDist pw) (hpz : forall w, IsDist (pz w)) (hpt : forall w z, IsDist (pt w z))
    (hu : forall t a, 0 <= u t a and u t a <= 1) :
    0 <= valueWZ pw pz pt u - valueW pw pz pt u and
      valueWZ pw pz pt u - valueW pw pz pt u <= sqrt (CMI pw pz pt / 2)

theorem conversation_boundR {M : Conv X A W Z} (hV : M.Valid)
    {beh : FedHist A W Z -> A -> Real} (hbeh : forall f, IsDist (beh f))
    (H : Nat) (hH : 1 <= H) :
    exists v, IsDeafValueR M H v and
      hearingValueR M beh H - v
        <= (1 + M.c * (H - 1)) * sum_{j < H-1} sqrt (iotaTopR M beh H (j+1) / 2) and
      sum_{j < H-1} sqrt (iotaTopR M beh H (j+1) / 2)
        <= sqrt ((H - 1) / 2 * sum_{j < H-1} iotaTopR M beh H (j+1))

theorem estimate_identity {pw pz pt} (hpz : forall w, IsDist (pz w))
    (hpt : forall w z, IsDist (pt w z)) (lw lj : Real) :
    estimate lw lj = CMI pw pz pt + exW pw pz pt lw - exWZ pw pz pt lj

theorem info_ge_corrected {pw pz pt} (hpz ...) (hpt ...) (lw lj : Real)
    (hex : 0 <= exWZ pw pz pt lj) :
    estimate lw lj - exW pw pz pt lw <= CMI pw pz pt

theorem corrected_exact {pw pz pt} (hpz ...) (hpt ...) (lw lj : Real)
    (hex : exWZ pw pz pt lj = 0) :
    estimate lw lj - exW pw pz pt lw = CMI pw pz pt

theorem CMI_le_of_factor [Fintype W] {A' : Type*} [Fintype A'] {pw : W -> Real}
    {pz : W -> Z -> Real} (pa : W -> Z -> A' -> Real) (pk : W -> A' -> Theta -> Real)
    (hpw : IsDist pw) (hpz : forall w, IsDist (pz w)) (hpa : forall w z, IsDist (pa w z))
    (hpk : forall w a, IsDist (pk w a)) :
    CMI pw pz (fun w z => push (pk w) (pa w z)) <= CMI pw pz pa

theorem vacuous_words_measurable {pw pz pt} (Hmarg lw : Real)
    (h : HW pw pz pt = Hmarg) : exW pw pz pt lw = lw - Hmarg
```

Where the formal statements differ from the text. The hidden state is one finite set X, with the type and the affect state folded into it; A, W and Z are finite and Z is nonempty. The formal Theorem 2 takes any behavioral hearing policy `beh` and exhibits a deaf value `v`, the value of the same policy fed fakes from a word-blind kernel, and bounds the gap to that value; the text's V_W^* is at least `v` because such a policy is a deaf policy, as the proof in Appendix A notes. The per-turn term `iotaTopR` is the survival-weighted conditional information under the running policy, as in the text. In Proposition 1, `estimate lw lj` is the fitted difference and `exW`, `exWZ` the two excess losses, so the identity is stated with the fitted losses as free reals and the excesses defined from them; `corrected_exact` requires the joint excess to be exactly zero, which is the text's "well specified", and `info_ge_corrected` takes the joint excess's nonnegativity as a hypothesis (`hex`) where the text derives it from Gibbs' inequality; `CMI_le_of_factor` states Theorem 1's second clause with the factoring written as a Markov kernel `pk` from the affect to the

type, and `vacuous_words_measurable` states Corollary 2's computable case as an identity in the marginal entropy.

Coverage. Every theorem, proposition and corollary of Section 2 and Appendix A is formalized, and the map above names each one's Lean statement. The build has no `sorry`, and the axiom check passes on all 33 checked theorems.

C Simulator and Protocol

Construction. The dialogue tasks exist as text (Liu et al., 2021; Stasaski et al., 2020; He et al., 2018; Wang et al., 2019). A planner that hears needs what none of them has: speech at both ends, affect in the delivery, and a hidden type to score against. We therefore draw a type, simulate the conversation, and render the user's turns as speech. The simulated user hears the system through an audio model that is not the system's encoder, so neither side can reward its own blind spot. Every component is public.

Types. Each task draws a hidden type θ. It carries the case, the goal and slot values that define success, the private parameters the task needs, whether the request lies inside the system's move set, which is false for a set fraction of cases so that fold has a correct use, and the affect dynamics, a transition model conditioned on the system's last act. The prior μ of Definition 1 is the sampling distribution of these types, so it is known exactly. Each affect state carries arousal, valence and dominance coordinates as well as a categorical name, so every turn has an exact label for both.

Speech. The user's utterances are rendered by `parler-tts/parler-tts-mini-v1`, prompted with a style instruction per affect label. A set fraction of turns is incongruent: the words are neutral while the voice carries the affect, so a planner reading the transcript alone is wrong on those turns by construction. Incongruence is drawn online per turn, so results are not split by it. The general-purpose encoder is `microsoft/wavlm-large`; the affect-tuned encoder is `audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim`, which also supplies the arousal-valence-dominance head. The simulator's four-dimensional channel is a projection of the affect-tuned encoder's pooled output onto the directions separating the affect classes. None of these models was trained for the simulator.

Calibration. The simulator is calibrated rather than assumed. Against ESConv (Liu et al., 2021) it reproduces the measured rates of disclosure, deflection and escalation at 0.130, 0.243 and 0.341, against 0.141, 0.273 and 0.322 in the corpus. A variant that leaks the hidden state into the words reaches 0.423, 0.666 and 0.807 and is rejected by the same comparison.

What each predictor sees. The words-only predictor is fitted on features of the turn's text. The joint predictor is fitted on those features together with the voice history, summarized as its mean over turns and its last value. The words-only predictor does not condition on earlier voices, so a per-turn estimate is a survival-weighted, clipped first difference of a cumulative curve rather than the increment ι_t of Theorem 2, which conditions on $z_{1:t-1}$. The voice enters through a two-point summary, so a channel whose information lives in its within-turn trajectory would be under-read.

Protocol. Runs use a horizon of $H = 8$, a turn cost $c = 0.02$, a fold payoff of $\rho_{\mathrm{in}} = 0.2$ when the goal was in reach after all, an incongruent rate of 0.35, and 300 paired episodes per arm at a fixed evaluation seed. The generator fixes the out-of-reach fraction at 0.4; the cue rate, the chance that a turn names the user's need in words, at 0.65; the affect-word rate at 0.2; the coupling of reachability to affect at 0.8; scope clarity, how far a named need settles whether it is in reach, at 0; and the near-clean channel's σ at 0.3. The exact planner draws 48 lookahead samples per act per decision, then re-scores the winning act on fresh samples, so that picking the best of several noisy estimates does not inflate its value. The hearing and deaf planners differ in one place, the last factor of equation 1, which is the encoder's likelihood for one and a constant for the other; users, seeds and the speech renderer are shared, and episodes are paired, with the pairing checked in code. The learned planner of Appendix D has one free parameter, the strength with which policy improvement is pulled back toward its starting point. It is selected once on held-out development episodes, disjoint from training

and evaluation, by mean return across both arms rather than by the gain, since selecting on the gain rewards a degraded deaf baseline, and it is held fixed across every condition.

D MEASUREMENTS IN DETAIL

This appendix follows the order of Section 3.

D.1 RESOLUTION FLOORS

The ceiling has a resolution, and its uncertainty has two sources rather than one. Resampling the evaluation episodes prices sampling variation but holds the fitted predictors fixed. The estimate is a difference of two fitted losses, so the fit moves as well, with its split assignment and its train-validation split. We measure both by repeating the whole estimate end to end across seeds. Under five-way cross-fitting at 300 paired episodes per arm the two are comparable: a resampling standard error of 0.0069 nats against a refitting standard error of 0.0067, with refitting accounting for 48 percent of the variance. The smallest ceiling separable from zero is then 0.027 nats for a single run of these arms, or 0.021 for the mean of five repeats. The two differ because every repeat scores the same episodes, so the resampling term is common to all of them and does not average down while the refitting term does. Neither is a constant of the estimator: across the ten hearing arms of the simulator, single-run floors span 0.022 to 0.048, a factor of two, and cross-fitted floors 0.017 to 0.042, the range the main text quotes, so each point is reported against its own floor rather than a shared one. The retention bounds above use the numerator's own floor at each coupling, 0.0103 to 0.0140 nats; those are floors of the binary in-reach estimand, a smaller quantity, and not this pooled figure.

Under a single train-validation split, which is the usual protocol, refitting is three quarters of the variance and the floor rises to 0.0865 nats. An oracle ceiling measured at 0.043 under that protocol is half its own floor, so it is not resolvable, and six repeats of it span 0.094. An interval from the episode bootstrap alone is too narrow by about a factor of $\sqrt{2}$ under cross-fitting and 2 under a single split, which is the difference between reporting a near-zero ceiling as resolved and reporting it as not resolved. Any measured ceiling below the floor for its protocol is reported as not resolved rather than as zero.

What sets the floor. Fitting the reported floors against a sweep over effect size and group count gives two exponents: the floor falls as the -0.56 power of the number of independent groups, and rises as the 0.36 power of the estimate itself. The relevant count is groups, speakers or dialogues, not items, which is why MELD's floor of 0.006 is the smallest reported here on $1{,}432$ dialogues even though its effect is also the smallest, while RAVDESS at 24 speakers carries the largest at 0.21. Neither term alone accounts for the spread.

Choosing the ridge. Both predictor classes select their ridge inside each training split, from a decade-spaced grid, on that split's own validation data, so the held-out split never informs the choice. Every estimate reported is therefore the best of a stated family rather than of one arbitrary penalty. We widened that family to eleven decades, 10^{-7} to 10^{3}, and re-estimated twenty-three rows on the same splits, seeds and bootstrap. No row changed its resolved-or-not verdict under either floor, the largest movement anywhere was 0.009 nats, and the synthesis comparison moved by less than a tenth of a percentage point on either side, leaving the ratios at 0.200 and 0.413. On MELD the joint class keeps selecting the same penalty on all twenty-five splits with three decades more on offer. The synthesized arm, which selects the largest value on the original grid, also keeps selecting it once ten, a hundred and a thousand are available, so that selection is interior rather than forced by the edge and the wider family reads 0.0013 nats *lower*, not higher. What moves under widening is the words side on the fixed-script corpora, where the words model is the class prior to within 0.004 nats, so its penalty is unidentified and its selection wanders four decades while the estimate does not follow. The side that carries the estimate is the side that stays put. On the simulator's null arms the selected joint ridge scatters across the whole grid, which is what selection looks like when the validation objective is flat because there is nothing to fit.

Table 1: The check on fixed-script corpora, split by speaker. D is the estimate with a 95 percent interval; loss_W and loss_{WZ} are the held-out log-losses of the words-only and the words-and-voice predictors; all are in nats. On a fixed script the words-only loss should equal the label's entropy; it does on RAVDESS and falls 0.005 nats below it on pooled CREMA-D. The shuffled control vanishes.

	D (nats)	loss_W	loss_{WZ}	Floor
CREMA-D	$+1.0583$ $[+1.0024, +1.1142]$	1.7853	0.7270	0.0799
CREMA-D, audio shuffled	-0.0001 $[-0.0004, +0.0001]$	1.7853	1.7855	0.0004
RAVDESS	$+1.1898$ $[+1.0453, +1.3343]$	2.0611	0.8713	0.2064
RAVDESS, audio shuffled	-0.0022 $[-0.0054, +0.0010]$	2.0611	2.0633	0.0046

D.2 Fixed-script corpora

Per-sentence estimates on the corpora. RAVDESS's two sentences are not prosodically equivalent (Major & Chatterjee, 2026), so we also estimate within each, where the transcript is constant by construction rather than merely uninformative on average. Those rows hold at 1.044 and 1.081, both excluding zero. On the synthesized side the same split gives 0.240 and 0.210. Splitting halves the sample, so the per-sentence floors are correspondingly higher and these rows corroborate the pooled estimate rather than adding precision to it.

D.3 MELD

Robustness of the MELD estimate. Two of the corpus's 13,708 utterances have no audio and are dropped. Three encoders and two joint classes, same splits, seeds, bootstrap and floors; the words-only loss is identical to the last digit across every audio representation (1.07254 under Bay, 1.12342 under sentence embeddings), which is the control that only the audio changed. Under the Bay words class the voice carries 0.196, 0.226, 0.286 and 0.270 nats through WavLM-large, the audeering affect model, Whisper large-v3 pooled over valid frames, and the same Whisper pass pooled over its padded thirty-second frame, and the words absorb 92.4, 90.8 $[88.4, 93.0]$, 91.9 $[89.8, 93.8]$ and 90.4 $[87.4, 93.1]$ percent of it, leaving 0.015, 0.021, 0.023 and 0.026; under sentence embeddings 89.7, 87.0, 87.4 and 83.8, the largest conditional term of the sweep being 0.044 nats, 2.8 percent of the label's entropy. Every conditional arm is resolved and every shuffled control is under its floor. Per nat of text-model improvement along the four words classes of Section 3.3, the estimate falls by 0.28 nats at the first step, then 0.10, then 0.07. A fourth words class, the seven held-out label scores of the transcript-only fine-tune of LLaMA-Omni 2, has a words-only loss of 0.998 and leaves 0.0099 $[0.0069, 0.0129]$ nats against a floor of 0.0044, with its shuffled control at -0.0004 and the unconditional arm reproduced to the last digit. Handing speaker identity to both predictors makes the words model slightly worse, because the transcript already carries it, and holding every test speaker out costs the voice arm 0.0019 nats, which is what speaker identity independently carries about the label. On the joint class, a 256-unit interaction network stacked on the split's own ridge term, the class that contains the published fit and starts from it exactly, moves the Bay term from 0.0150 to 0.0158 and the sentence-embedding term from 0.0203 to 0.0210; a plain network in the joint slot reads below the ridge on every arm, the unconditional included, so it is a different family with its own optimizer and not a superset.

D.4 Off-the-shelf models

Scoring the off-the-shelf models. Each model receives the utterance's transcript under a fixed instruction, in two phrasings, once without the audio and once with it, through its own chat template; the two conditions differ only in the presence of the audio. Its distribution over the seven labels is read without generating: each label string is scored teacher-forced after the prompt, and the seven string log-probabilities are normalized over the seven. The two sets of label probabilities then enter the estimator in the words slot and the voice slot, beside the corpus's own words features, so the model's calibration is fitted by the estimator rather than assumed, and the reported share uses the same splits, seeds, bootstrap and floors as the ceiling.

Models and prompts. QWEN2-AUDIO is `Qwen/Qwen2-Audio-7B-Instruct`, run in bfloat16, and LLAMA-OMNI 2 is `ICTNLP/LLaMA-Omni2-7B`; the resolved model revisions are recorded in the released results. Both models receive the same two instructions, with the utterance's transcript in place of $\langle line \rangle$; in the audio condition the model's chat template inserts the audio block before the same text. No token is generated.

> *Plain.* Which emotion does the speaker express in this line? Line: "$\langle line \rangle$" Answer with exactly one word from this list: anger, disgust, fear, joy, neutral, sadness, surprise.
>
> *Framed.* The following is one line spoken by a character in a scene from a television sitcom. The character says: "$\langle line \rangle$" Classify the emotion the character is expressing as exactly one of: anger, disgust, fear, joy, neutral, sadness, surprise. Reply with the single label word and nothing else.

What two off-the-shelf models collect. The ceiling they are measured against is read through WavLM, so it is a reference rather than a bound for models that hear through their own encoders. We scored two speech language models on the same 13,706 utterances, the same dialogue splits, the same five seeds and the same words class, routing their own label scores through Proposition 1's estimator rather than differencing their raw losses. Against the 0.0203 ceiling that words class gives, QWEN2-AUDIO collects 36.2 percent $[24.6, 47.9]$ under one prompt and 31.5 $[20.5, 42.4]$ under the other, both resolved against their own floors. LLAMA-OMNI 2 collects 2.2 percent $[-1.3, 5.6]$ and 1.7 $[-0.9, 4.4]$, neither resolved. Every shuffled-audio arm sits under its own floor. So one model reaches about a third of the headroom and the other reaches nothing we can measure, on a quantity whose whole range is one percent of the label's entropy.

The naive measurement disagrees with all of this by two orders of magnitude. Differencing the transcript arm against the audio arm gives $+0.508$ and $+0.510$ nats for the first model and $+0.053$ for the second, which are 69, 80 and 151 times the disciplined figures; audio drawn from a *different* utterance reproduces 74, 72 and 66 percent of them. Their transcript arms sit 1.3 to 1.9 nats above the words probe, and Proposition 1 inflates a difference between two suboptimal predictors by exactly that excess, so a zero-shot arm this far from optimal can produce a raw difference of any sign without contradicting Theorem 1; the second model's is negative under one of its two prompts. The mechanism is visible in the label mix. MELD is 47 percent neutral, and audio moves mass toward the majority label whether or not it is the matching audio, which buys log loss with no affect perceived. The two models differ in their input path only in scale, so this is evidence about what such a difference measures, not about either architecture. Fine-tuning LLAMA-OMNI 2 on the task, a low-rank adapter on the language model with the speech encoder and projector frozen, over the same five splits, gives an audio-and-transcript model and a transcript-only control with held-out losses of 0.9982 and 0.9981 nats; the raw difference is -0.0001 $[-0.0032, +0.0030]$, the disciplined gain on the sentence-embedding class is $+0.0025$ $[-0.0016, +0.0067]$ against a floor of 0.0127, and the shuffled-audio arm reads $+0.0022$ against a floor of 0.0118. Fine-tuning fixed the model's label prior and made it the strongest text model here; it did not make it hear. Training the audio pathway does not either. With the speech projector unfrozen alongside the adapter, 20.5M more parameters over the same splits, the audio arm's loss rises to 1.00116 against the transcript-only control's 0.99807, a raw gain of -0.0031 $[-0.0053, -0.0005]$; every disciplined arm is unresolved ($+0.0015$ against a floor of 0.0092 on the sentence-embedding class, -0.0001 on Bay), and shuffled audio reproduces 114 percent of the raw difference.

D.5 SYNTHESIS AND THE PLANNER'S CHANNEL

Synthesis. On the three emotions COSYVOICE3 documents instructions for, split by recording, the human row reaches 0.895 $[0.842, 0.949]$ nats and the synthesizer under those instructions 0.853 $[0.808, 0.899]$, 3.8 points below on the three-label scale and inside the interval; under the English instructions we wrote it reaches 0.509, and PARLER-TTS 0.062. Our own label-to-style bridge for PARLER-TTS mapped *neutral* and *calm* to one string, so two of its six classes were the identical input; on the four labels with no tie the gap between the two synthesizers narrows and does not close, 25.9 against 43.2 percent. The synthesized conditions are split by recording because a synthesizer has no speakers, and the human condition is split the same way, which lets its predictor learn each actor. Held out by speaker instead, the actors' six-label estimate is 0.941 nats against a floor of 0.216, 53.2 percent of the label's entropy; against it the two synthesizers keep 29 and 59 percent of what the actors carry, where the matched split gives 20 and 41.

The encoder and its projection. A ceiling is worth reporting only if something can approach it. A calibrated head on the same pooled output, the same recordings and the same speaker-disjoint splits reaches 0.736 accuracy against a 0.171 chance rate and a cross-fitted calibrated log loss of 0.739 against the marginal's 1.790. The ceiling run's own predictor reaches 0.740, so a plain classifier lands within two hundredths of a nat of the estimate: what binds is the representation, not the head that reads it. Fine-tuning the encoder end to end on the same splits (a head warm-up, then the full trunk at a learning rate of 10^{-5}, the only one of three recipes that was stable) does not change that: its own head reaches 0.681, below the frozen probe, while the estimator's converged readout on its features reaches 0.756 and the ceiling through them is 1.092 $[1.042, 1.143]$ nats against the frozen 1.058 $[1.002, 1.114]$, under both floors. Its four-dimensional class-mean projection carries 0.684 $[0.636, 0.733]$ against the frozen projection's 0.357, and its six logits 1.020; the shuffled-audio control through it reads -0.0002.

D.6 SIMULATOR

Table 2: What the policy collects, as the channel improves. α is the probability that a turn is observed through a near-clean reading of the affect state, a one-hot vector in Gaussian noise at $\sigma = 0.3$, rather than through the encoder, so $\alpha = 0$ is the encoder alone on synthesized speech. *Gain* is the hearing arm's mean return minus the shared deaf arm's, in the return units of Definition 1. $\hat{\iota}$ is the unclipped log-loss difference averaged over the arm's turns, not summed over them, so it is an estimate of a nonnegative quantity and can come out negative when that quantity is small; each is the mean of five end-to-end repeats, with 95 percent intervals carrying resampling and refitting. All five points share one deaf control and one population.

α	Estimated $\hat{\iota}$ (nats)	Gain
0.00	-0.0035 [-0.0117, $+0.0047$]	$+0.004$ [-0.066, $+0.076$]
0.25	-0.0029 [-0.0140, $+0.0081$]	$+0.144$ [$+0.077$, $+0.209$]
0.50	$+0.0134$ [-0.0162, $+0.0430$]	$+0.237$ [$+0.177$, $+0.300$]
0.75	$+0.0110$ [-0.0157, $+0.0376$]	$+0.273$ [$+0.215$, $+0.332$]
1.00	$+0.0320$ [$+0.0052$, $+0.0589$]	$+0.266$ [$+0.208$, $+0.326$]

Reading Tables 2 and 3. In this simulator the words carry nothing about reachability, so the voice is the only route to it. Sweeping the coupling moves the amount sent by a factor of five while the amount arriving stays at zero, so the flat *Kept* column is a channel erasing what is sent rather than a generator sending nothing. The gains are direct measurements. The $\hat{\iota}$ column is not: no arm clears its own single-run floor, and because the arms do not hold the words channel fixed, a better channel folds sooner and degrades the words predictor along the same axis, the direction Proposition 1 warns of, so that column is an upper bound. The *Kept* column's intervals are Wald, the *Sent* column's a percentile bootstrap, and every ceiling interval carries both resampling and refitting.

Shuffled inputs and the learned planner. Shuffling the audio against the transcripts drives the estimate to zero: the estimator invents nothing. Shuffling the *policy's* input answers a different question. A hearing arm fed an affect drawn independently of the user, matched in scale, gains $+0.006$ $[-0.064, +0.074]$ beside the real encoder's $+0.004$ and cuts over-extension from 0.662 to 0.233

Table 3: What the channel delivers. *Coupling* sets how strongly the speaker's reachability drives her affect state; 0.80 is calibrated against ESCONV and is the operating point used throughout. θ_{reach} is the in-reach component of the hidden type, *Sent* is measured at the generator before any channel, *Kept* is the ceiling after synthesis and encoding, and brackets are 95 percent intervals over episodes and refits. *Retention* is a bound, not an estimate, because *Kept* covers zero everywhere: it is that column's own resolution floor over *Sent*.

Coupling	Sent, $I(\theta_{\text{reach}}; e)$	Kept, $I(\theta_{\text{reach}}; Z \mid W)$	Retention
0.10	0.0518 [0.0249, 0.1074]	$+0.0010$ [-0.0089, $+0.0108$]	< 0.270
0.30	0.0704 [0.0353, 0.1369]	-0.0056 [-0.0139, $+0.0027$]	< 0.168
0.50	0.1477 [0.1001, 0.2089]	$+0.0043$ [-0.0048, $+0.0134$]	< 0.088
0.80	0.2635 [0.2228, 0.3134]	$+0.0030$ [-0.0056, $+0.0116$]	< 0.047
0.95	0.2724 [0.2355, 0.3144]	$+0.0017$ [-0.0055, $+0.0089$]	< 0.038

against the real encoder's 0.271, the two hearing arms differing by -0.002 $[-0.067, +0.065]$. The behavioral signatures the channel appears to produce therefore reproduce when the channel carries nothing: the planner's Gaussian likelihood moves its belief more than the four numbers' information warrants, so the signatures measure the belief update rather than the information. A planner trained by anchored rollout policy iteration reproduces them as well: fold rate 0.080 deaf, 0.333 on the real channel and 0.340 on a shuffled one, over-extension 0.850, 0.571 and 0.579, and a difference of $+0.007$ $[-0.037, +0.050]$ between its two hearing arms. It gains $+0.019$ $[-0.033, +0.071]$ against the exact planner's $+0.004$, indistinguishable at $+0.015$ $[-0.059, +0.089]$; its approximation error is the size of the quantity being measured, which is why the instrument is an exact solver.

Power of the bound's estimator. Table 4 subsamples the 300 paired episodes of each arm without replacement, 200 draws per cell with a fresh refit seed each, and compares the bound's estimator $B = R\sum_t \sqrt{\hat{\iota}_t^+/2}$ with the 95th percentile of its own centered null. It reaches eighty percent power near 200 episodes at $\alpha = 0.75$ and 160 at $\alpha = 1$; the unclipped sum, which is unbiased under the null and not a bound, behaves alike, so the clip costs little power. At $\alpha = 0$ the estimator rejects in 7.5 percent of draws.

Table 4: Power of the Theorem 2 bound estimator against its centered null, by episodes n.

n	25	50	75	100	150	200	250	300
$\alpha = 0.75$	0.03	0.10	0.26	0.26	0.60	0.79	0.89	0.98
$\alpha = 1.0$	0.06	0.16	0.45	0.58	0.78	0.92	0.97	1.00

The bound's estimator near zero. The estimator of the bound is biased upward near zero. Both clips in ι_t are one-sided, so a turn carrying no information contributes the positive part of its noise, and the square root rewards many small positive terms; centering the per-turn differences so that every true increment is exactly zero, then resampling, gives the distribution the estimator produces under a true null. With $H = 8$ and $c = 0.02$, $R = 1.140$, and the bound $R\sum_t \sqrt{\hat{\iota}_t^+/2}$, formed from per-turn clipped estimates rather than from the averaged $\hat{\iota}$ of Table 2, reads 0.187 to 0.503 across the sweep. It is not monotone in α, because the number of turns carrying an estimable increment falls as the channel improves and conversations end sooner. The measured gain sits under it at all five points, and against the largest gain, 0.273, it is about twice the measured effect; at $\alpha = 0$ it reads 0.187 against a null of 0.136, so there most of it is the estimator's own noise. This is a scale check rather than a test of Theorem 2, whose ι_t the estimates lower-bound while the bound's estimator is biased upward. Subtracting the null does not repair the bias, which shrinks as the true information grows: a null measured at zero over-corrects wherever the channel is good, so the corrected quantity can fall below the truth and is no longer a bound. A correction valid across the range needs the bias estimated at the observed ι_t rather than at zero.

E Extended Related Work

Table 5: What each bound covers. The first two rows are Wei (2024)'s closest results: the value of one perfect observation for one decision, and an open-loop planner's gap from the optimal policy. With payoffs in $[0, 1]$ his $R_{\max}$ is 1, so the first comparison is like for like; ι_t is the per-turn information of Theorem 2. Appendix E explains why his Theorem 5.5 is not a competitor.

	What is bounded	Constant	Holds for
Wei (2024), Prop. A.8	one perfect observation, one decision	$R_{\max}\sqrt{2I}$	one Bayes decision
Wei (2024), Thm. 5.5	open-loop planning, against optimal	linear in I	$I \geq 2$ nats per step
Theorem 1	hearing over the words alone	$\sqrt{I/2}$, sharp	one Bayes decision
Theorem 2	hearing over the best deaf policy	$R\sqrt{\sum_t \iota_t/2}$	any running policy

Value of information. Ordering channels by decision value is Blackwell's (Blackwell, 1953), and Proposition A.8(iii) is his garbling theorem on the fold region; the exact gap between two channels

Table 6: Measurements of whether a spoken system hears: what each measures, its control, and what it is scored against.

Work	Quantity	Control	Scored against
Foo et al. (2026)	benchmark score with and without audio	none	label
Xiong et al. (2026); Lu et al. (2026)	accuracy on stimuli built to conflict	the conflict itself	label
Koduru et al. (2026); Pang et al. (2026)	probes of encoder layers	none	label
Liu et al. (2026)	optimal-action rate with audio, with the true state, and with a written-down reading	oracle state	decision
Rakib et al. (2026); Wang et al. (2026a)	task success	none	language-model critic on the transcript
This paper	log-loss gap with the Proposition 1 audit; per-turn ceiling of Theorem 2	shuffled audio; a resolution floor per run	hidden type in the simulator, label on the corpora

is Le Cam's deficiency (Le Cam, 1964), which Theorem 1 trades for something estimable; the value of an experiment is Lindley (1956). Theorem 1 is the decoupling inequality of information-ratio analyses (Russo & Van Roy, 2016; Russo & Zou, 2020; Xu & Raginsky, 2017) with Pinsker's sharp constant (Kullback, 1967); the many-turn version for log utility is a directed information (Permuter et al., 2011) and Theorem 2 is its bounded-payoff counterpart. Table 5 lists Wei (2024)'s Theorem 5.5 to say that it is not a competitor for Theorem 2: it bounds an open-loop *planner*'s gap from the Bayes optimal policy, where both policies still observe and update their beliefs, rather than a hearing policy's advantage over a deaf one, and its information term enters linearly because the assumption $I \geq 2$ nats converts $\sqrt{2I}$ into I. That assumption cannot hold in a conversation where each turn carries a little, and we know of no prior bound of Theorem 2's form. Affect in a dialogue belief goes back to Bui et al. (2009). The log-loss difference is the established estimator of the prosody-information literature (Wolf et al., 2023; Yadavalli et al., 2026); what we add is that it bounds what a policy can gain, and Corollary 2 for when it does not. Proposition 1 is the log-loss form of excess risk: a misspecified conditional model's excess loss is its expected divergence from the true conditional (Cover & Thomas, 2006). The beliefs at which folding is optimal form a convex set that shrinks as the voice improves and grows as the horizon runs out (Proposition A.8), the shape known from sequential decision problems (Arrow et al., 1949; Smallwood & Sondik, 1973). It is a conditional $\mathcal{V}$-information (Xu et al., 2020; Hewitt et al., 2021), estimable for the same reason it is one-sided, which is what the negative results on estimating information from samples (Antos et al., 1999; McAllester & Stratos, 2020) say. The question is the expected value of sample information (Howard, 1966; Raiffa & Schlaifer, 1961; Frankel & Kamenica, 2019; Sims, 2003; Caplin & Dean, 2015), computed here from the two channels a spoken system already records.

Games of incomplete information. The system faces a user whose type is private (Harsanyi, 1967); conversation itself has been modeled as strategic choice of words in pragmatics (Parikh, 1991; Asher & Lascarides, 2013). Words are cheap talk, and cheap talk with conflicting interests admits an uninformative equilibrium (Crawford & Sobel, 1982); a channel the sender does not fully control can inform where words cannot (Spence, 1973), and the voice is such a channel, since affect leaks through it (Ekman & Friesen, 1969). The voice is harder to control than the words (Zuckerman et al., 1981), carries affect in pitch, rate and voice quality (Scherer, 2003; Laukka et al., 2011; Sporer & Schwandt, 2006), is read by listeners about as well as faces (Juslin & Laukka, 2003), and can nonetheless be acted (Scherer, 2013; Jürgens et al., 2011). Bargaining with one-sided private information and an outside option (Rubinstein, 1985; Fudenberg & Tirole, 1983) is the negotiation task; the impossibility of efficient bilateral trade under private values (Myerson & Satterthwaite, 1983) bounds what any policy can achieve there. The stopping act has cousins: the reject option in classification (Chow, 1970), deferral to an expert (Madras et al., 2018; Mozannar & Sontag, 2020), and the outside option in bargaining (Binmore et al., 1989). From computational game solving we take two tools: scoring leaves over a belief rather than a point estimate (Brown et al., 2018; Moravčík et al., 2017) and improving a policy under a regularizer to a human-like prior (Meta Fundamental AI

Research Diplomacy Team (FAIR) et al., 2022; Bakhtin et al., 2023). The closest models give the receiver a signal about the sender that the sender does not control. With such a signal, cheap talk can be effective under conflicting interests (Seidmann, 1990), and it can also be crowded out, so that the receiver would rather not have the signal (Lai, 2014; Ishida & Shimizu, 2016; Moreno de Barreda, 2024); Proposition A.9 is that effect in our game. When the signal reports whether the words are true, it disciplines talk (Kartik, 2009; Dziuda & Salas, 2018; Balbuzanov, 2019; Sadakane & Tam, 2026). In bargaining with a public news channel and a cost of delay, the uninformed party's signal has positive value while her offers add nothing in equilibrium (Daley & Green, 2020), the closest dynamic analogue of fold.

Dialogue policy planning. Proactive planners choose a dialogue act each turn on emotional support, tutoring, negotiation, and persuasion tasks (Liu et al., 2021; Stasaski et al., 2020; He et al., 2018; Wang et al., 2019). The current design proposes candidate acts with a frozen language model, selects one with a small value network trained online, and tracks the user's emotion from text in the state and the reward (Rakib et al., 2026); rollout planning with user portraits reports full success on the same benchmarks (Wang et al., 2026a). Rollouts for negotiation planning go back to Lewis et al. (2017) and Yarats & Lewis (2018), and belief-state planning for dialogue to the partially observed formulation of Young et al. (2013). Every planner above judges success with a language-model critic reading the transcript, and none hears the user. We define success against the user's hidden type, add deferral as an act, and put audio-derived affect in the state.

Spoken interaction and affect. Modular spoken assistants stream speech from a language model's text (Fang et al., 2025a;b; Wang et al., 2025) and native models carry affect implicitly (Défossez et al., 2024); the former cannot condition output style on the input's paralinguistics, by their own account. Reward-driven spoken systems align to paralinguistics from a judge (wen Yang et al., 2026), learn turn-taking rewards (Chen et al., 2026), or decouple when to speak from what to say (Li et al., 2026; Hsiao et al., 2026); none plans over acts with a hidden type. Audio language models keep most of their scores without audio and follow the transcript when it contradicts the voice (Foo et al., 2026; Pang et al., 2026; Huang et al., 2026; Wang et al., 2026b), which is why we evaluate where words and voice disagree. The closest work traces that failure to task decisions: Liu et al. (2026) hold a task fixed, hide the user's concern, and compare four conditions. Direct audio access barely improves on the transcript and a true-state oracle gains twenty-six points, but a model that infers the concern from audio into an explicit text representation before acting recovers all but one of them, so the information is present and the architecture decides whether it reaches the decision.