Beyond Text

The Text-Centric Bias in Foundation Models Must Be Revisited for a Speech-First Future

Deepak Babu Piskala

Microsoft, Seattle, WA, USA

ICML 2026 · Position Paper Track Spotlight

Seoul, South Korea · July 2026

Interactive version of the paper:

prdeepakbabu.github.io/speech_icml_paper

The Position in One Slide

TL;DR

Text-centric foundation models reflect **interface habit**, not necessity.

As voice becomes habitual, training data will become **speech-first** motivating **speech-native model architectures**.

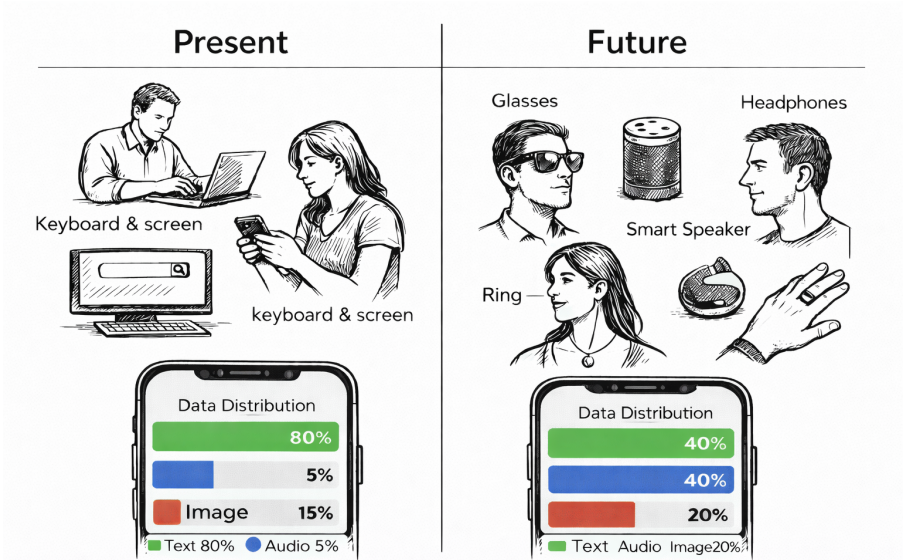
What we argue

- Speech tech is now **ready** (Whisper, GPT-4o)
- Adoption gap is **habit**, not capability
- Data ecosystem will **shift toward speech**
- Text-pretrained backbones will misalign

What we don't argue

- **Not** that speech replaces text
- **Not** that today's compute math is wrong
- **Not** that adoption flips overnight
- **Not** a specific architecture prescription

The Shift Underway



Why Does Text Dominate AI?

The intuitive answer:

"Text is the natural way to express ideas."

Why Does Text Dominate AI?

The intuitive answer:

"Text is the natural way to express ideas."

The historical answer:

Text dominates because:

- Keyboards came before microphones
- ASR was unreliable for decades
- Search boxes shaped query habits
- Interfaces conditioned expression

Why Does Text Dominate AI?

The intuitive answer:

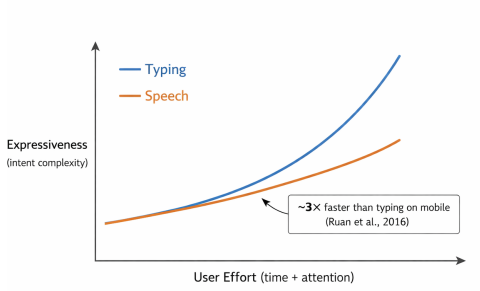
"Text is the natural way to express ideas."

The historical answer:

Text dominates because:

- Keyboards came before microphones
- ASR was unreliable for decades
- Search boxes shaped query habits
- Interfaces conditioned expression

Text feels natural because it is habitual.



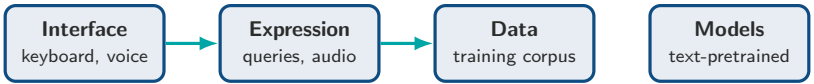
Speech is $\sim 3\times$ faster than typing on mobile.
Ruan et al., 2016

The Interface Feedback Loop

The Interface Feedback Loop

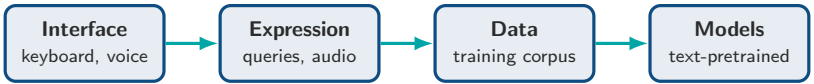
> **Interface shapes expression** search boxes condition keyword lists; voice enables natural questions

The Interface Feedback Loop



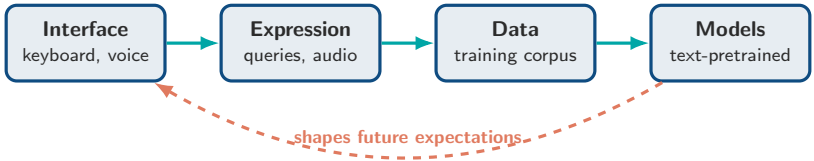
- **Interface shapes expression** search boxes condition keyword lists; voice enables natural questions
- **Expression shapes data** what users externalize becomes the corpus

The Interface Feedback Loop



- **Interface shapes expression** search boxes condition keyword lists; voice enables natural questions
- **Expression shapes data** what users externalize becomes the corpus
- **Data shapes models** modalities and biases in data become inductive biases in models

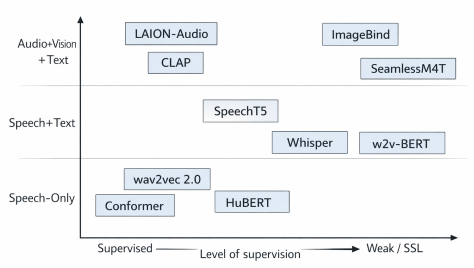
The Interface Feedback Loop



- **Interface shapes expression** search boxes condition keyword lists; voice enables natural questions
- **Expression shapes data** what users externalize becomes the corpus
- **Data shapes models** modalities and biases in data become inductive biases in models
- **Models then shape what users expect interfaces to be** closing the loop

Speech Tech Has Crossed the Usability Threshold

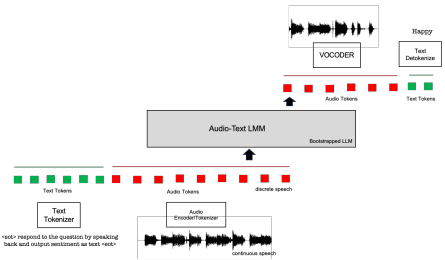
- **Whisper** (680K hrs, 100+ languages)
  robust ASR across accents
- **HuBERT, wav2vec 2.0**
  self-supervised audio representations
- **GPT-4o**
  232ms response latency (human range)
- **Gemini Live**
  bidirectional streaming with barge-in



From supervised speech-only to self-supervised multimodal.

The remaining gap is not modeling capability.

Speech-Native vs. Speech-Grafted



Audio-LLM stack: most systems today bootstrap from a text LLM.

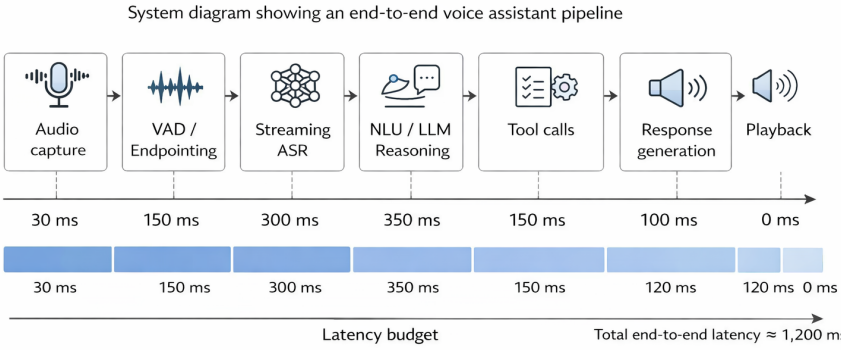
Today (grafted)

- Text LLM backbone
  PaLM, LLaMA, Qwen
- Audio discretized to fit text-style tokens
- Paralinguistic signal often collapses to text

Tomorrow (native)

- Audio-first pretraining
- Semantic audio tokenization
- Preserves prosody, emotion, identity

The Latency Gap (Closing Fast)



Traditional cascade

ASR → LLM → TTS

∼1,200 ms

Speech-native (GPT-4o)

end-to-end audio

∼232 ms (human norm: 200 ms)

Habit Inertia: The Residual Barrier

Once technical barriers fall, what remains is **accumulated behavioral and institutional infrastructure** built around text.

The feedback loop is the cause, not just the effect:

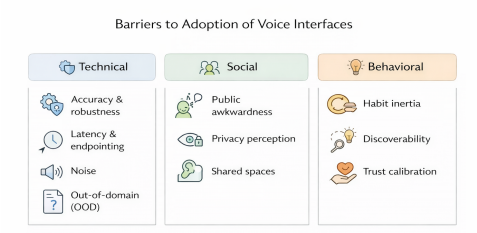
- Users were trained on keyboards because that's what worked
- Habits shaped the data
- Data shaped the models
- Models now reinforce the text default

Habit Inertia: The Residual Barrier

Once technical barriers fall, what remains is **accumulated behavioral and institutional infrastructure** built around text.

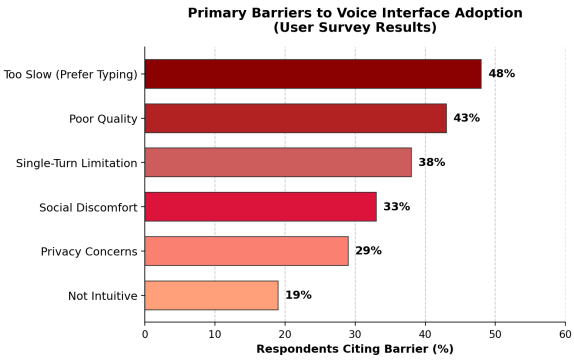
The feedback loop is the cause, not just the effect:

- Users were trained on keyboards because that's what worked
- Habits shaped the data
- Data shaped the models
- Models now reinforce the text default



Habit is itself a product of prior tech constraints.

Survey Evidence (N=200, Illustrative)



Key signals:

- **76%** use voice at least occasionally
- **100%** use voice at home
- **Only 25%** use voice at work
- Top barriers: *speed*, *quality*, *single-turn*, *social*, *privacy*

Caveat: informal online poll, self-selection bias toward tech-literate respondents. Presented as illustrative pattern, not population study.

Alternative Views (Addressed in Paper)

"Text is more efficient"
Audio is 10–100× more expensive per FLOP.
Response: cost is a research target, not a permanent veto. Hierarchical tokenization closing the gap.

"Paralinguistic = noise"
Tone, emotion are extra cost for marginal value.
Response: healthcare, education, customer service, safety-critical systems depend on it.

"Text has real advantages"
Searchable, editable, private, precise.
Response: agreed. Text wins for search, formal docs, non-linear editing, and accessibility. Both modalities deserve first-class treatment.

"Habit shift won't happen"
QWERTY persisted; voice will too.
Response: marginal data shift is already underway. Voice search exceeds 20% of mobile queries; voice messaging dominates in WhatsApp and WeChat.

Call to Action: Three Concrete Directions

1. Audio-first pretraining

Train foundation models on audio from the start.

Open Q: Does it produce different internal phonological structure? Different scaling laws? More robust to accents and code-switching?

2. Semantic audio tokens

Move beyond phonetic clustering (HuBERT, wav2vec).

Goal: hierarchical tokens that preserve content, prosody, and speaker characteristics at distinct levels.

3. Eval w/o text intermediation

Build benchmarks measuring reasoning *over* audio.

Goal: audio-native counterparts to MMLU, HumanEval.
ProfASR-Bench is a template.

Research questions, not solved problems. The paper's job is to argue they deserve priority.

Privacy: A First-Order Architectural Constraint

Not a deployment afterthought. A pretraining-time design constraint.

Why speech is different:

- Voiceprints (biometric)
- Ambient sounds (context leak)
- Continuous capture (no clear boundary)
- Emotional state (sensitive signal)

What the agenda must include:

- Consent-aware data pipelines
- On-device processing for sensitive audio
- Federated learning adapted for speech
- Differential privacy for continuous signals

GPT-4o system card already treats voice ID + generation as safety risks.
Community should formalize this.

Multilingual Robustness & Accent Equity

A core blocker, not future work.

ASR accuracy drops substantially for:

- Low-resource languages
- Non-standard accents and dialects
- Code-switched speech

These populations are often **the most speech-first** in daily life: limited literacy infrastructure, oral knowledge traditions.

The speech-native argument is both more urgent and harder there.

Treat as core, alongside

- Scale
- Efficiency
- Privacy
- **Multilingual robustness**
- **Accent equity**

Not after architectures stabilize on English.

For Industry & the Broader Community

Industry practitioners:

- Invest in speech data infrastructure (podcasts, meetings = tomorrow's training data)
- Deploy speech-native interfaces (generates training signal, shifts habits)
- Open-source speech-native models (GPT-4o, Gemini Live are proprietary)

Community:

- Build audio-native benchmarks (beyond GLUE/MMLU)
- Engage HCI & cognitive science (habit barriers need interdisciplinary work)

Takeaways

1. **Text dominance reflects interface habit, not cognitive necessity.**
   Decades of keyboards conditioned how users externalize knowledge.
2. **Technical barriers to voice have largely fallen.**
   Whisper, GPT-4o, Gemini Live demonstrate readiness.
3. **What remains is habit inertia + architectural inertia.**
   Speech-language models still initialize from text backbones.
4. **Future training data will trend speech-first.**
   Voice search, voice messaging, podcasts, meetings, wearables.
5. **The community should anticipate this, not react to it.**
   Audio-first pretraining · semantic tokens · text-free evaluation · privacy · multilingual equity.

Thank you.

Questions?

Position: Beyond Text
ICML 2026 · Position Paper Track Spotlight

Deepak Babu Piskala · prdeepak.babu@gmail.com