Company

Introducing Accent Conversion for the listener

Mar 03, 2026

Written by Arto Minasyan

The scale of the problem

Why training a ML model for accent understanding Is a Harder Problem Than You Think

Krisp Voice AI Lab

Try it today — and build with us toward what's coming

I have a PhD in Mathematics. I've built Krisp — a global Voice AI company used by millions. We've built 8 voice technologies that process sound at the edge in real time. I negotiate term sheets with tier-1 Silicon Valley VCs. On paper, I'm a reasonably intelligent person.

Then I open my mouth on a Zoom call. "Sorry, could you repeat that?" "I think you're breaking up." "Can you maybe... type it in the chat?"

PGR2026-00032
Krisp Technologies, Inc. v. Sanas.AI Inc.
Patent Owner Exhibit 2028

It's a strange experience — knowing you can solve differential equations and DSP problems but watching someone's face glaze over because you pronounced "model" in a way their brain didn't expect. You go from "builder of 8 voice AI technologies" to the guy who repeated his order in a coffee shop 4 times.

For years, I assumed this was a me problem. My English isn't good enough. I need to practice more. I should watch more American TV shows. I should slow down. I should take accent coaching.
Then I did what any self-respecting engineer would do — I looked at the data. There are 1.5 billion non-native English speakers in the global workforce. That's more than the native speakers. The majority of English spoken on business calls today is spoken with an accent. We are not the edge case. We are the default.

And that's when it hit me: my accent isn't a personal failure. It's a signal processing problem. The gap isn't between my brain and my mouth — it's between my mouth and your ear, at least on a Zoom call. And that gap? That's an engineering problem. And we've spent years building the world's first accent understanding technology.

Let's start with a number that surprises most people: there are 1.5 billion non-native English speakers in the global workforce. Native speakers? About 400 million. Non-native speakers outnumber native ones 4:1. The majority of English spoken in business today is accented English. It's not the exception — it's the statistical norm. English is the most-studied language on Duolingo, ranking #1 in 154 countries, and it's the top language to learn in 79% of countries

Yet every piece of communication infrastructure built in the world— from Zoom to Google Meet to Microsoft Teams — is optimized for the minority case.

This plays out in two massive use cases.

Global teams on calls, all day, every day

Think about what happens when someone says "Can you repeat that?" on a Zoom call. It feels like nothing. Three seconds, maybe five. Now multiply that by the hundreds of millions of virtual meetings happening daily across global companies. Multiply it by the meetings where someone *didn't* ask to repeat — they just nodded, pretended to understand, and moved on with the wrong information.

a prospect missed a key point, in engineering specs that get misinterpreted, in decisions that take three meetings instead of one. It's a massive, invisible tax on global productivity, and nobody's measuring it because we've all just accepted it as friction.

Learning — where the stakes are even higher

Some of the most brilliant professors and educators in the world have strong accents. So do some of the best content creators on YouTube. The knowledge in their heads is world-class. But a meaningful percentage of their audience is only absorbing 70–80% of what they're saying — not because the content is hard, but because their listeners' brains are spending cycles decoding pronunciation instead of processing ideas.

This isn't just intuition — the neuroscience backs it up. Researchers at Washington University found that when listeners encounter unfamiliar speech patterns, their brains recruit extra cognitive resources just to map sounds to words — the same machinery that kicks in when you're trying to hear someone in a noisy bar. A 2025 study went further and measured it physiologically: listeners' pupils dilate measurably more when processing non-

comprehension budget.

But here's the part that should make you uncomfortable. That cognitive load doesn't just reduce understanding — it reduces perceived credibility. Researchers at the University of Chicago found that identical factual statements were rated as less truthful when spoken with a foreign accent — even when listeners were explicitly told the speaker was just reading someone else's words. The same sentence. The same facts. Rated as less true, simply because the listener's brain had to work harder to process it. And our brains, it turns out, interpret processing difficulty as a signal of unreliability.

This has been replicated consistently: accented speakers are rated as less intelligent, less competent, and less employable across professions and cultures. Not because of what they're saying. Because of how it sounds.

The same words. The same ideas. Degraded in transit. That's not a people problem. That's a channel problem. And broken channels are what engineers fix.

understanding Is a Harder Problem Than You Think

Most people hear “Accent AI” and assume it’s a straightforward fine-tuning job. Feed the model more accented speech, adjust the weights, ship it. We thought so too, briefly, at the beginning. Then reality arrived.

Here’s what actually makes this hard.

There’s no ground truth

The foundational requirement for supervised learning is simple: for every input, you need a labeled output to train against. For accent understanding, that label is a parallel recording — the same voice, the same words, the same prosody, but in a different accent. Imagine a dataset where every Indian-accented English speaker also has a matching recording of themselves speaking in Neutral American. Same person, same sentence, just a different accent.

That dataset does not exist. It has never existed. You can’t hire annotators to create it, because no annotator can make you sound like yourself with a different accent.

absence of this parallel data isn't a gap you can paper over with clever augmentation — it's a fundamental constraint that forces you to rethink the entire training paradigm from the ground up.

The accent space is essentially infinite

Even if you solve the labeling problem, you're facing a combinatorial nightmare. There are roughly 7,000 languages in the world. Each produces its own interference pattern when its speakers acquire English — different phoneme inventories, different prosodic structures, different vowel spaces. Then layer on regional dialects within those languages, urban vs. rural variation, age, education, code-switching. Two speakers from the same city, same age, same native language will sound meaningfully different.

You cannot enumerate the accents. You have to build a model that generalizes across a space it has never seen the edges of, and that generalization has to hold in production, in real-time, on a call where the stakes are a sales deal or a medical consultation.

Accent is woven into identity — the voice itself

timbre: the unique resonance of your vocal tract that no one else shares. Your pitch and its natural range. Your rhythm and cadence — how you pause, how you breathe, how you land on certain words. Your prosody — the melody of your speech. Your emotional texture. And your accent — the phoneme patterns, vowel shapes, and consonant placements that your native language carved into your English over years of use. All of these are deeply entangled in every millisecond of audio you produce.

Our goal is to reach into that bundle, isolate the accent dimension, soften it or convert it to Neutral American, and put everything else back exactly as it was — so that when you hear the output, you still recognize the speaker. Same timbre. Same rhythm. Same person. Just more intelligible. Disentangling one dimension of identity from all the others, modifying it, and reconstructing the full signal without touching anything else is an extraordinarily hard representation learning problem. The model has to learn what makes your voice *yours* and what makes it *accented* — and those two things are not neatly separated in the data. They never are.

Generating high-quality voice with a tiny model

parameters. ElevenLabs, Voicebox, the frontier systems: they're large because high-fidelity speech generation is hard. Every nuance of prosody, formant transition, and breath pattern that your brain uses to assess authenticity requires modeling capacity.

We're doing this at the edge. On-device. With a model small enough to run on consumer hardware without cooking the CPU. That means we had to rethink the architecture entirely rather than just compress a large model down — compression loses exactly the high-frequency detail that separates natural speech from uncanny speech. The engineering challenge is generating quality that passes the ear test using a fraction of the compute that the industry assumes is the minimum.

Real-time or it doesn't exist

A model that makes you more intelligible after a 2-second delay is not a communication tool — it's a liability. On a live call, latency above roughly 250ms breaks the conversational loop. People talk over each other, responses feel disconnected, the interaction degrades in ways that are worse than just having the accent.

produce an output. The model has to make high-quality predictions with partial context and do it quickly. If the model produces word artifacts, it might make one think they are speaking with a bot. If the model uses a heavy amount of computation, this may delay output production, resulting in information loss in the audio, something one may observe during connection issues. , The real-time constraint doesn't just change the speed requirement — it changes the fundamental architecture of what's buildable.

On-device: the final boss

All of the above, and it has to run locally. No cloud round-trip. The reasons are obvious once you think about it: privacy, latency (cloud adds milliseconds you can't afford), and reliability (calls drop, VPNs throttle, hotel WiFi is a disaster). On-device is the only deployment model that works in the real world for this use case.

But on-device means the model's parameter budget is measured in megabytes, not gigabytes. It means running on CPUs with wildly varying capabilities. It means the model that works perfectly on an M3 MacBook has to also work on a three-year-old Windows laptop running twelve other applications. The optimization surface is

required

Krisp's accent understanding works on the listener side, at the audio driver level, below any application. It creates a virtual speaker that sits between whatever conferencing software is running and your physical output device. The incoming audio — the voice of your Indian colleague, your Ukrainian founder, your Filipino support agent — gets processed and transformed to US native in that layer, in real time, on your device, before it reaches your ears.

Zoom, Teams, Meet, any proprietary dialer — none of them need to know Krisp exists. They just see a speaker. This matters because the problem is universal across every platform and every call. A solution that only works inside one app isn't a solution. Krisp processes every incoming voice at the output layer, and every platform upstream feeds into it automatically. When a new conferencing tool launches tomorrow, accent understanding works with it on day one. No update required. No integration required. No permission required.

We didn't fully appreciate how hard this was when we started. Years later, we have a clearer view: accent understanding sits at the intersection of representation learning, real-time audioprocessing, on-device

approached the problem. There's also a reason we kept going.

Krisp Voice AI Lab

Eight years ago, we set out with a mission that sounded ambitious to the point of absurdity: build the most critical real-time voice AI technologies in the world — and run them entirely on-device.

We started with noise cancellation. Then background voice cancellation. Then we kept going: accent localization, accent understanding, real-time speech-to-text — all processing audio at the edge, on your device, with no round trip to the cloud. In 2025, we expanded into server-side technology and shipped real-time voice translation supporting 63 languages — one of the best in class. Now we're scaling that server-side stack further, building out STT and TTS to complement our on-device foundation.

If you want to see the full picture of what we've built and what we're working on, visit https://lab.krisp.ai.

Eachof these technologies was built in Yerevan, Armenia.

engineers for decades — what it hasn't had is a company that gave that talent a stage to compete at the global frontier. That's what Krisp is.

There's something specific that happens when a team of researchers from a small country decides they're going to build technology that outperforms what comes out of Silicon Valley, London, or Beijing. There's no safety net of brand recognition. No default assumption of credibility. You ship something that works, or you don't matter. That pressure produces a particular kind of engineer — one who is resourceful, rigorous, and slightly obsessed.

Every technology listed above — noise cancellation used by millions, real-time translation, accent understanding — was built by people who know exactly what it feels like to be underestimated because of where they're from. That's not incidental to the work. It's fuel for it.

Try it today — and build with us toward what's coming

today and your entire team gets clearer on every call, across every platform, without changing a single tool in your stack. If you're a developer building communication infrastructure, the Krisp SDK exposes the same capability directly: accent understanding you can embed into your own product, your own pipeline, your own platform.

One thing worth saying plainly: this is beta. The technology works, and the effect is real — but we're at the beginning of what's possible. We've been building this for three years and we know exactly where the edges are. Over the next year, as the models mature and the training data compounds, the quality will improve substantially. Accents that are harder today will get easier. Edge cases will close. The gap between what we're shipping now and what we believe is achievable is large — and that's not a caveat, it's the reason we're excited.

The problem has existed for decades. The infrastructure to solve it is only now becoming viable. If you're on global calls every day and you've normalized the friction, you don't have to.

Start the discussion...

LOG IN WITH

OR SIGN UP WITH DISQUS

Name

Share

Best Newest Oldest

Related Articles

Voice AI:...

May 6, 2026

Company
Engineering Blog SDK

—...

April 1, 2026

Company Krisp News

for Customer...

February 18, 2026

Company
Engineering Blog SDK

Krisp Appoints CX Industry Veteran Harry Folloder to...

February 12, 2026

Company Enterprise
Krisp News

Introducing Krisp's Accent Conversion SDK

December 16, 2025

Company
Engineering Blog SDK

Introducing 3.5x smaller Voice Isolation model with exceptional...

December 2, 2025

Company
Engineering Blog
Enterprise SDK

Audio-Only Turn-Taking Model v2

November 3, 2025

Company
Engineering Blog SDK

Krisp Launches Accent Conversion for Africa

October 16, 2025

AI Accent Conversion
Company
Contact Centers
Enterprise Krisp News

Introducing Krisp Companion Mode: Stay Present in Every...

October 8, 2025

Company Product