RUTGERS
WINLAB | Wireless Information Network Laboratory

Ellison Murray[1] **Alexander Chen**[2] **Anya Knudsen**[3] **Adam Jean**[4] **Morriel Kasher**[1] **Predrag Spasojevic**[1]

[1] Rutgers University [2] Purdue University [3] Cornell University [4] Carnegie Mellon University

Granulated Low-bitrate Mixture

An open-source lightweight audio codec for speech recognition on tiny devices

3.01% word error rate

24 kbps bitrate

4.16× less encoder work

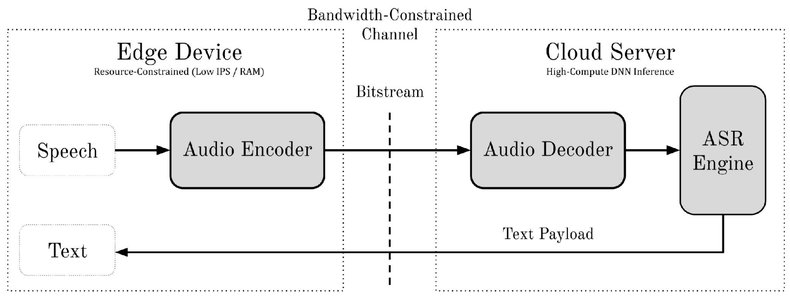
0 floating-point ops

The problem

Speech recognition is everywhere, but the smallest devices (earbuds, smart glasses) have kilobytes of memory and no floating-point unit.

They cannot run the model. The alternative is to offload: a tiny encoder on the device, the heavy model on a server.

But raw audio drains the battery and the bandwidth, so it must be compressed first, and the compression itself has to be nearly free.



The split we target: a small encoder on the device, the heavy recognition model in the cloud.

Everything the device runs is that one encoder box. So encoder cost and decoder quality is what we minimize.

What we optimize for

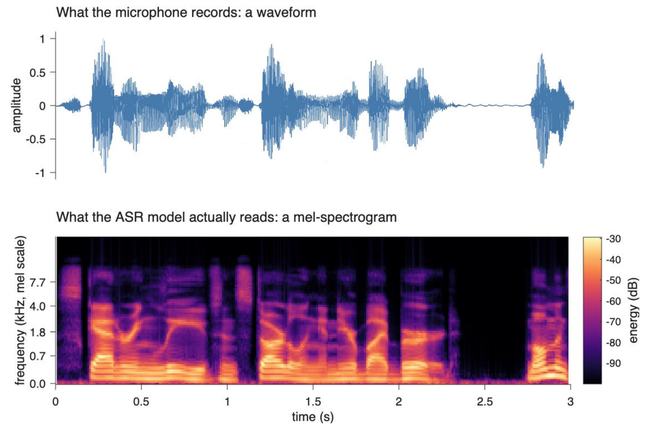
Instructions
how much math per second of audio
battery life

Memory
how much RAM the encoder holds
does it even fit?

Bitrate
bits per second over the link
bandwidth cost

And what we deliberately do NOT optimize for:
how good it sounds. MP3 and Opus are tuned for a human ear. No human ever hears our audio: only a machine does.

What the machine hears



Top: the waveform a microphone records. Bottom: the same sound as a mel-spectrogram. Illustrative synthetic speech.

Models never read the waveform. They turn it into a picture: x axis is time, y axis is frequency, brightness as energy.

The model runs image-style convolutions over it. Whatever distortion is added to that picture, the model treats as speech.

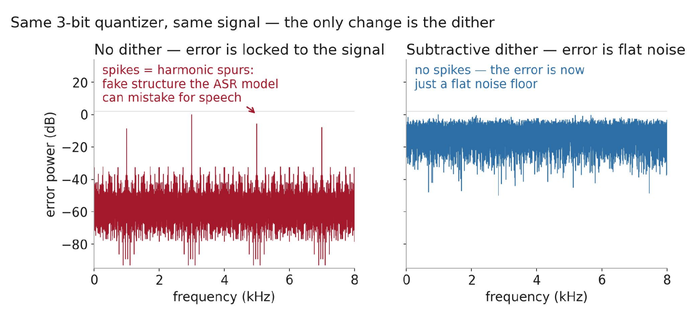
The trick: subtractive dithering

1 Add
the encoder adds a small random value v to each sample

2 Round
quantize as usual — the rounding is now randomized

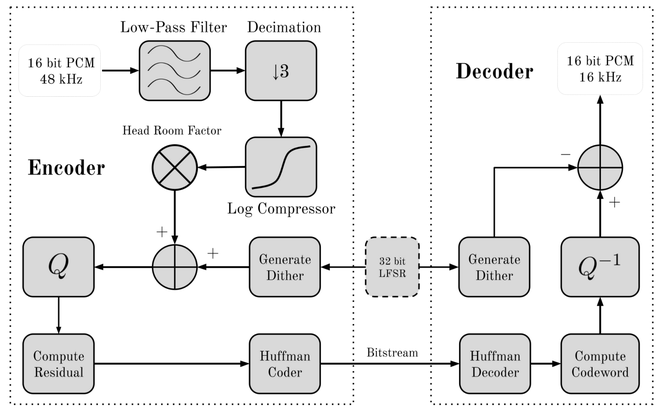
3 Subtract
the decoder removes the exact same v, regenerated from a shared seed

Randomizing the rounding breaks the link between signal and error. The error stops being structured and becomes ordinary hiss.

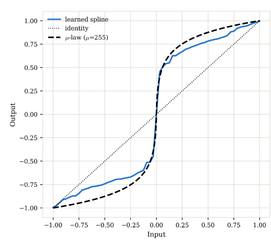


Same 3-bit quantizer and signal — only the dither changes. Left: harmonic spurs the model can mistake for speech. Right: a flat noise floor.

What we built

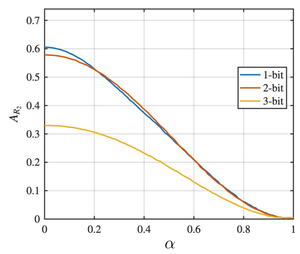


The GLX encoder and decoder — integer arithmetic only, no floating-point operations anywhere.



A learned compression curve

Trained to preserve information, it rediscovered μ-law — used by telephone networks for decades.



One dial, α, for how much dither

Turning α up removes the fake structure but costs bitrate. We swept it and picked the best point per bit depth.

Results

Codec	Encoder instr. (M)	kbps
G.711	0.345	64.0
GLX	2.25	23.7
G.726	9.66	24.0
MP3	16.9	24.0
SILK	24.2	23.4

Millions of instructions per second of audio, lower is better. Only G.711 is cheaper — and it needs 64 kbps against our 26. Evaluated on 2 620 held-out LibriSpeech utterances with Whisper large-v3.

The honest trade: at matched bitrate our accuracy trails MP3 and SILK, and we are more sensitive to background noise. On a chip with no floating-point unit and a battery to protect, that is the right side of the trade off.

WINLAB