babblelabs

Extreme Neural Network Computing Transforms Speech Quality

Chris Rowen

CEO

BabbleLabs Inc.

The speech opportunity today

8B

Phones + radios + TV + PCs delivering voice

13B

Minutes per year of YouTube Uploads

20B

Microphones today

200T

Minutes per year of device interaction

1Q

Words per year in voice calls

7.6B

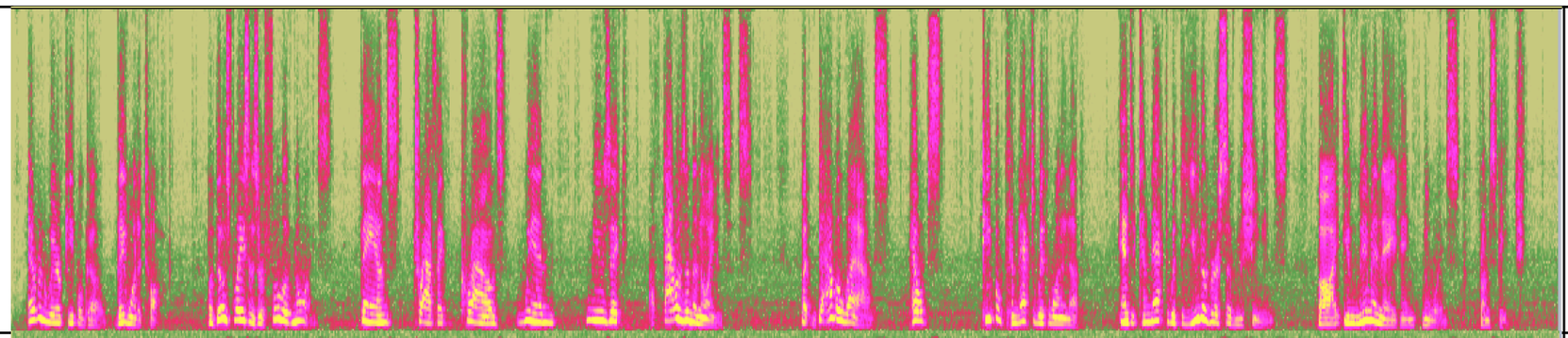
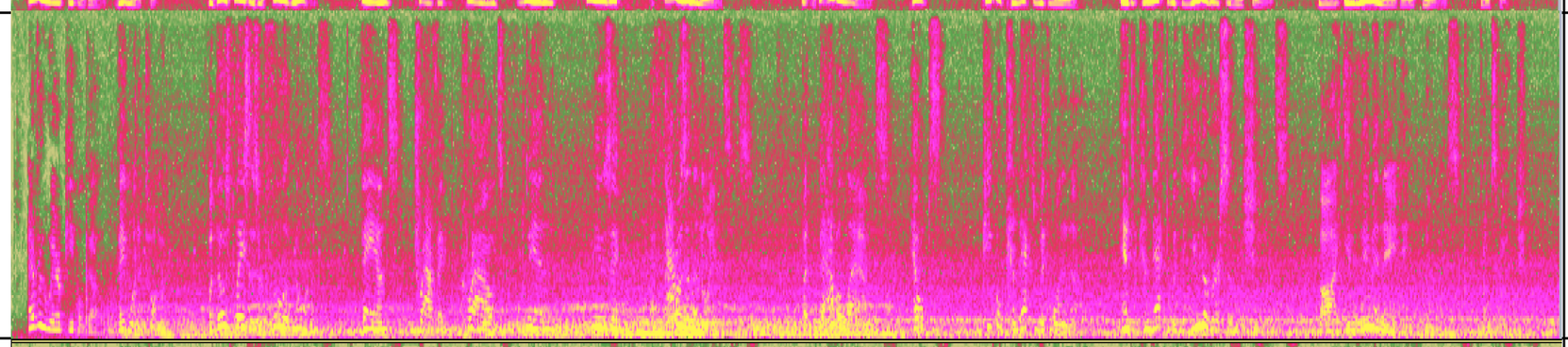
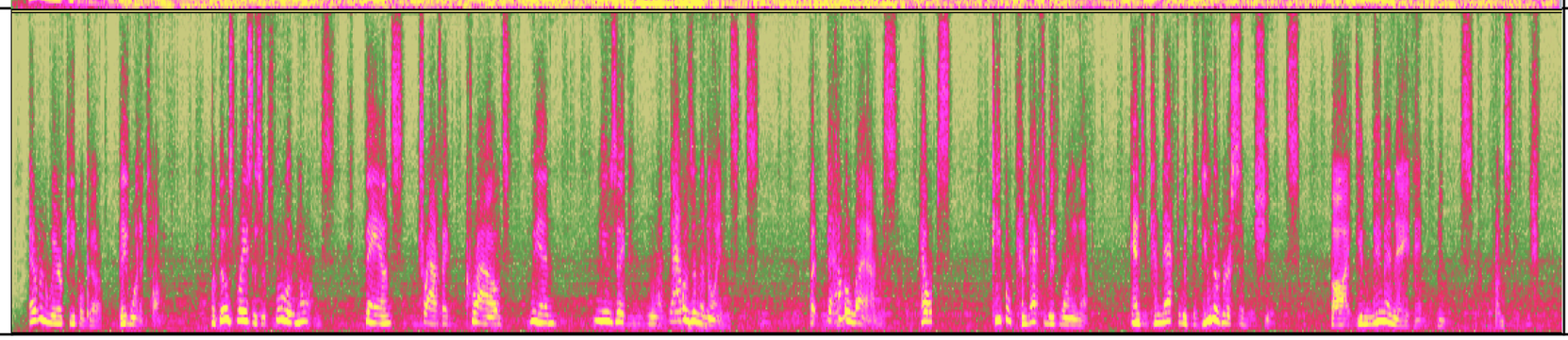
People

But often too frustrating to understand and use in the real world

babblelabs

The big problem: noise

The answer: speech enhancement

Clean speech		
Speech with car noise (0dB)		
Enhanced speech		

Once you remove the noise, then you can do much better on …

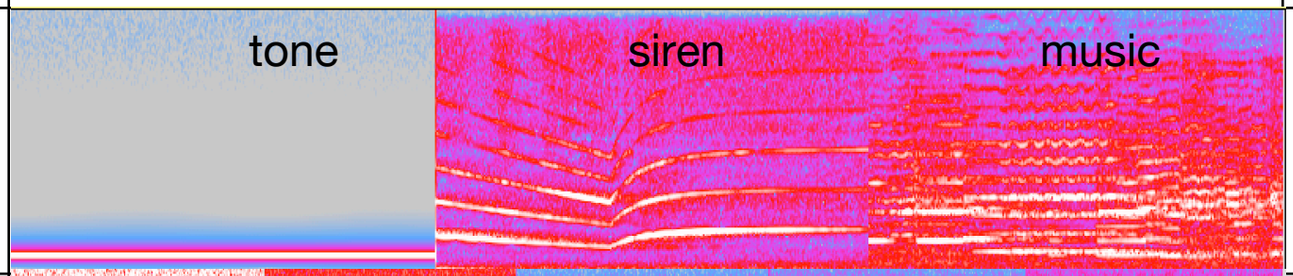
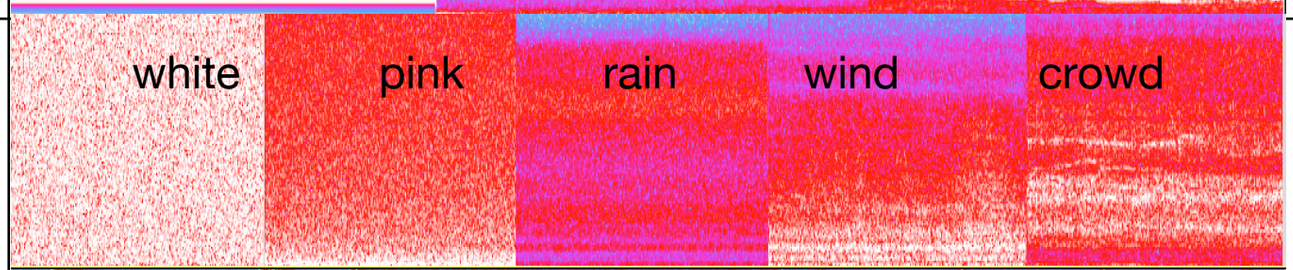
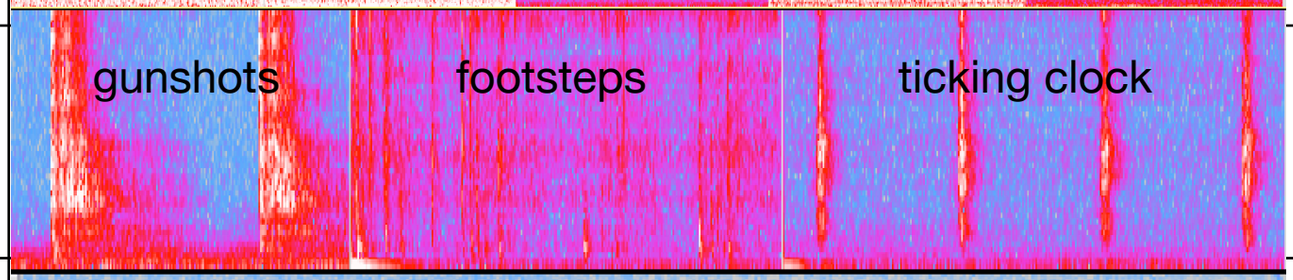
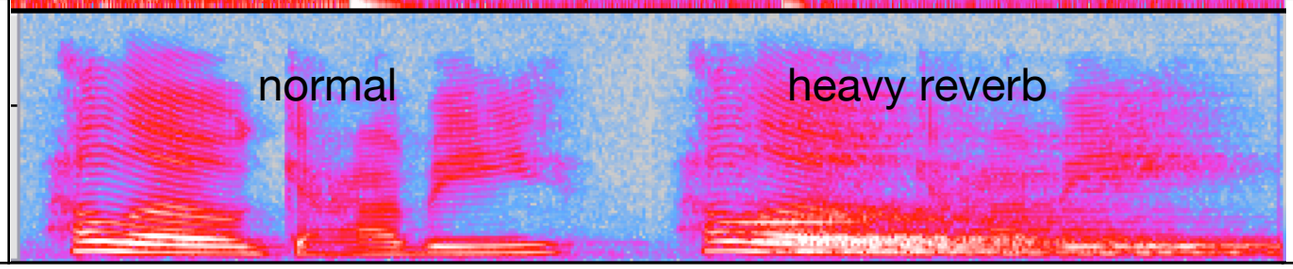
…audio and video recording

…live phone calls and video chat

…speech recognition

babblelabs

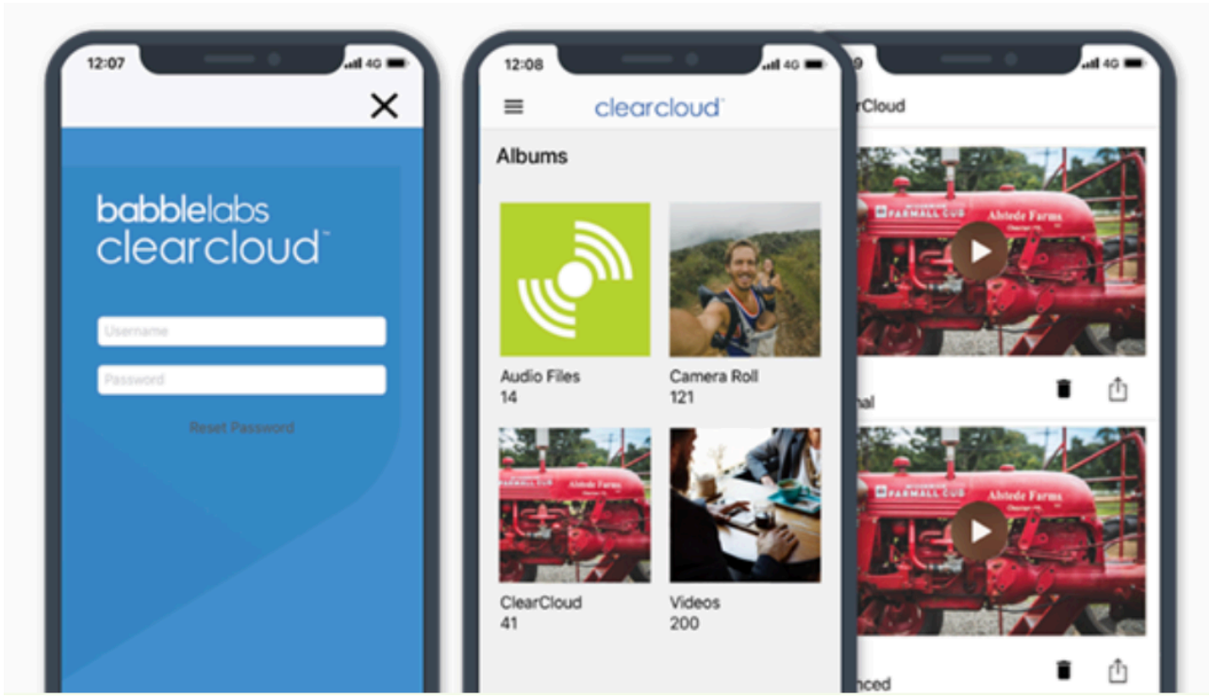
A Field Guide to Noise

Noise Type	Spectrogram	Audio	What to look for
Narrow band			• Tones and music • Pronounced frequencies • Stationary vs. non-stationary
Wide band			• Wind, fans, static • Overlaps with and destroys speech • Background speech (pureéd)
Transitory noise			• Bangs, claps, gun shots, and sound effects • Highly non-stationary
Reverberation			• Added echoes of once clean speech • “Smears” speech into a muddle

babblelabs

The Cloud

- A simple API:
  - `output_stream` ←**`/audioEnhancer/api/{audio,video}/stream`**
    `(email, token, stereo_out_flag, input_type, input_stream)`
  - `output_file` ←**`/audioEnhancer/{audio,video}/uploadFile`**
    `(email, token, stereo_out_flag, input_type, output_type, input_file)`
  - `token` ←**`/accounts/api/auth/login`**
    `(email, password)`
- Current audio formats: aac, aiff, mp4, m4a, mpeg, ogg, wav
- Current video formats: mp4, mov
- Deployed as web-service and iPhone and Android audio/video capture apps:



babblelabs

Try it yourself – free BabbleLabs audio/video apps

iPhone

Android

App never stores your data in cloud or shares it with BabbleLabs

babblelabs

Going Native

- Cloud deployment works for some use-cases, but limitations abound:
  - Latency
  - Computing cost and power
  - Questions on data privacy
  - Network availability
- Neural network inference getting dramatically easier at the edge:
  - Embedded GPU
  - PC clients
  - NN accelerators in phone Aps [originally intended for vision]
  - Embedded DSPs
  - Microcontrollers
- High-performance computing engines + algorithmic ingenuity + implementation ingenuity yields undetectable latencies in the telephony stack.

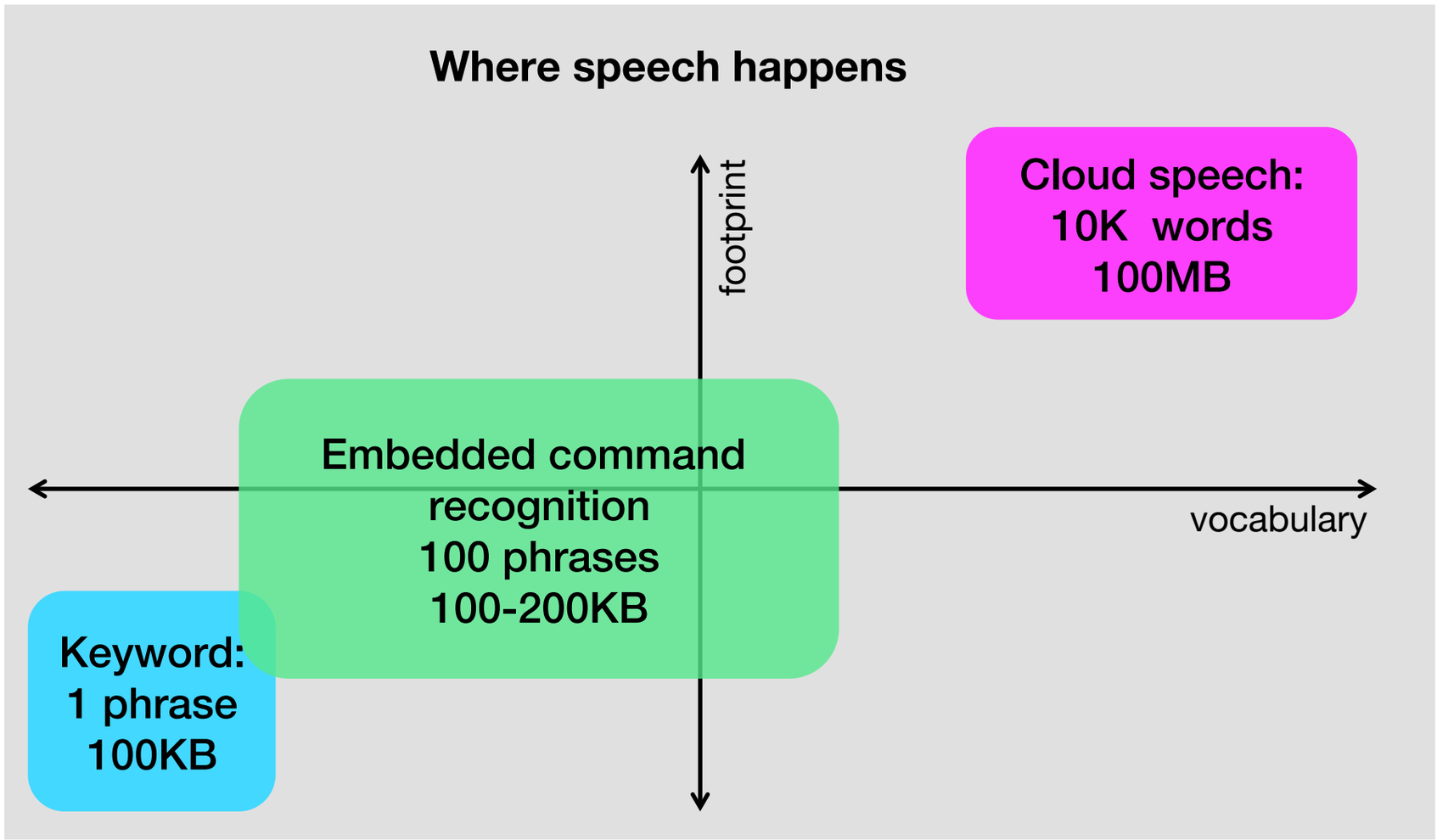
➔A new era of near-perfect speech in telephony

BabbleLabs Clear Cloud

Deep learning breakthrough for speech enhancement: ***demo video***

Make ASR Better

BabbleLabs Clear Command



babblelabs

Make ASR Better

- Traditional ASR struggles with noise:
- Word recognition unconstrained by expected vocabulary
- Waveform-to-meaning combines phoneme/word extraction, language model, and phrase recognition

Command ID	Example: 80 phrases for 35 commands		
0	turn on the TV	turn on the television	
1	turn off the TV	turn off the television	
2	turn up the TV	turn up the television	
3	turn down the TV	turn down the television	
4	turn on the AC	turn on the air conditioner	turn on the air conditioning
5	turn off the AC	turn off the air conditioner	turn off the air conditioning
6	turn up the AC	turn up the air conditioner	turn up the air conditioning
7	turn down the AC	turn down the air conditioner	turn down the air conditioning
8	turn on the lights		
9	turn off the lights		
10	turn up the lights		
11	turn down the lights		
12	turn on music	turn on the music	turn on the sound
13	turn off the music	turn off music	turn off the sound
14	turn up music	turn up the music	turn up the sound
15	turn down music	turn down the music	turn down the sound
16	turn on the heat		
17	turn off the heat		
18	turn up the heat		
19	turn down the heat		
20	open menu	open the menu	show the menu
21	open music	show music	
22	open maps	show maps	
23	open Facebook	show Facebook	
24	open Twitter	show Twitter	
25	open Instagram	show Instagram	
26	open browser	open a browser	open the browser
27	open weather	show weather	
28	open messages	show messages	
29	open photos		
30	open WeChat	show WeChat	
31	what time is it?	what's the time?	
32	what's the weather?		
33	answer the phone	answer phone	answer telephone
34	show the news	open the news	show news

The Training Challenge

- High quality output ➔

sophisticated neural models ➔

huge training sets ➔

intense training

Training data

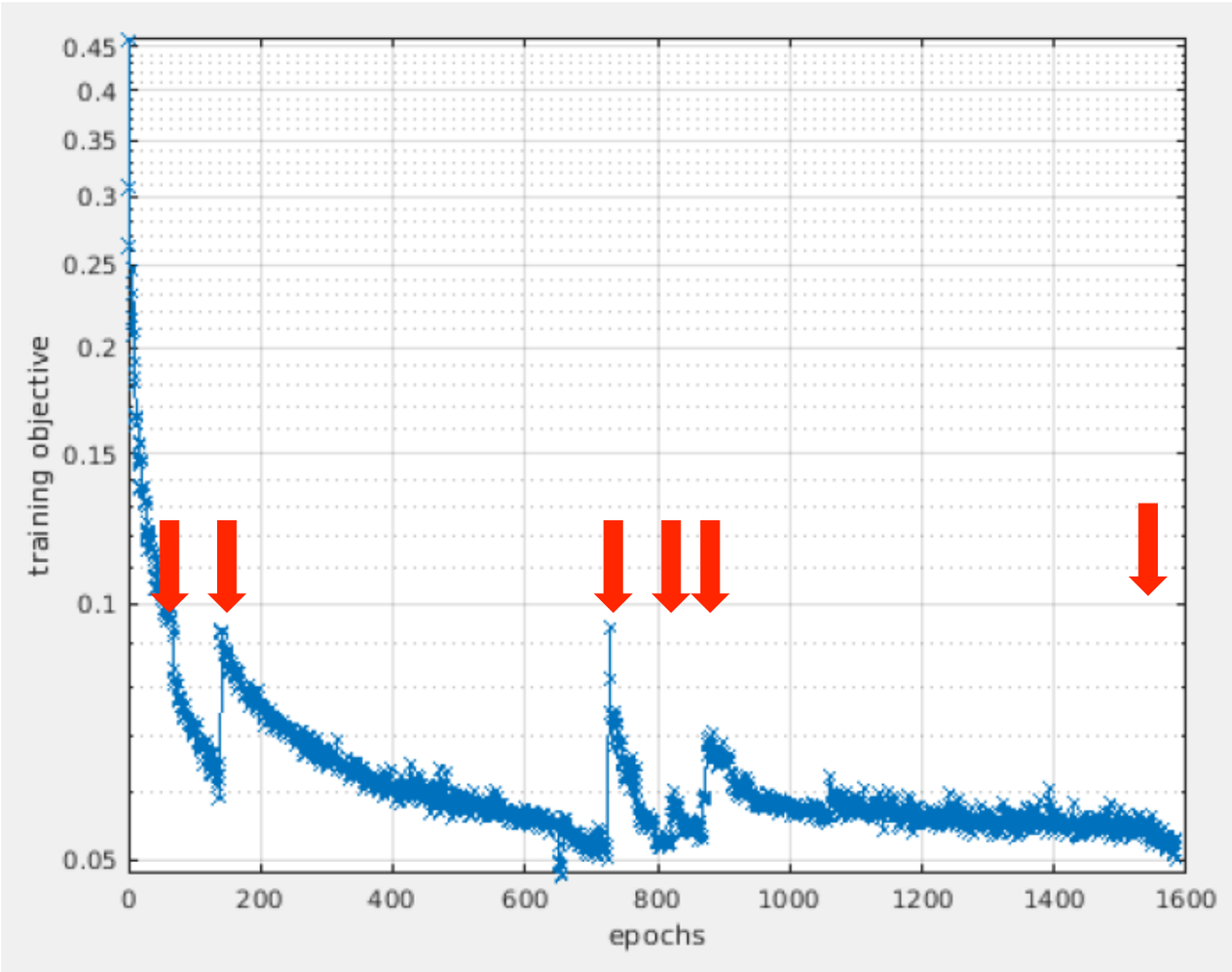
- Unique collection + augmentation of noise, speech, room models
- Raw corpus:
  - 40,000 hours speech
  - 15,000 hours music
  - 15,000 hours noise
  - 100,000 room acoustic models
- Typical training set size: ~10M minutes of noisy speech

Full training time:

1,000 hours x 8 NVIDIA V100

On-the-fly updates:

- added training data
- enhanced loss functions
- model branching for domain-specific variants



Change in learning rate or loss function, or added noisy speech content

babblelabs

Mixed public cloud and in-house GPU environment

3 PetaFLOPs for development

- Four NVIDIA compute cluster types:
  - In-house GPU workstations: 1080ti
  - Public cloud:
    - training experiments: V100 or P100
    - production model training: typically 4-8 V100 cluster
    - production API service: optimized GPU or CPU code:

      Top choices: K80, P100, P4, T4, x86
- Distributed computation:
  - many parallel experimental and release-candidate trainings
  - data augmentation servers + training servers

GPU Type	GPU Usage	Peak TFLOPs
P100	41	870
1080ti	20	230
V100	16	1,920
Total	**77**	**3,020**

The BabbleLabs System

Live Network Engines

Clear Edge

Clear Mobile
Speech Enhancement Network

Clear Command
Speech Recognition Network

Clear Cloud

Clear Cloud
Speech Enhancement Network

Standard Cloud Speech Recognizers

Backend Services

Model Server

Network Training

Network Customizer

Speech Analytics

Speech Corpus

Noise Corpus

Acoustics Corpus

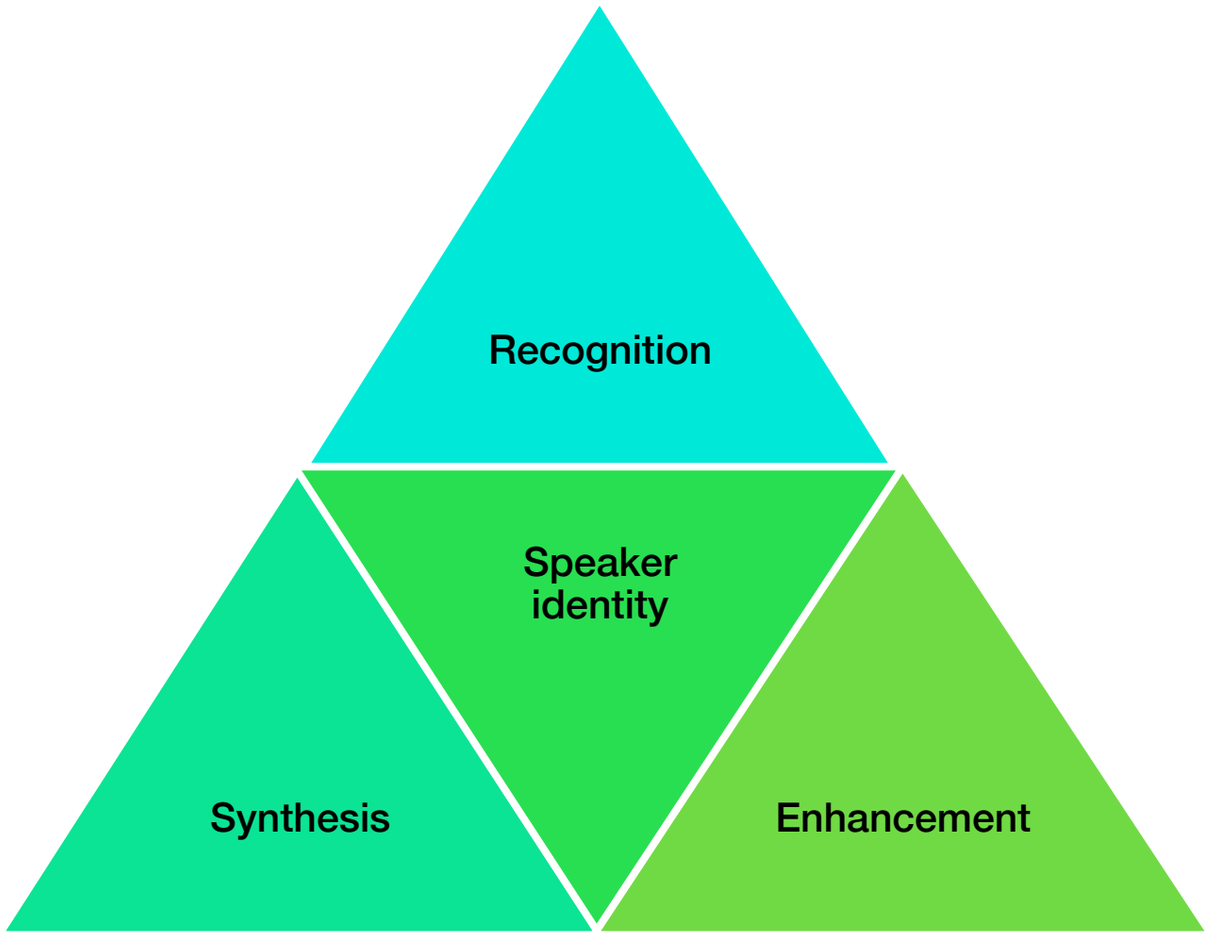
Use-Case Segments

Telephony | Media | Gaming | Education | Customer Service | Public Safety | Car | Home/IoT

babblelabs

The future of speech: Speech >> Text

- Speech is much more than a live text stream: emotion, sentiment, health, environment, ID
- Remarkable progress to date on speech clarity and speech recognition
- Noise remains huge issue for real-world proliferation of smart speech
- Solving the “cocktail party problem” - de-muxing competing speakers – is within reach
- Speech systems interact – anything you learn about speakers and noise in one function helps other functions perform better
- Speech IS the preferred UI for people – machines will finally adapt to us!



babblelabs

speak your mind