Empowering Customer-Facing Teams with Voice-Based AI

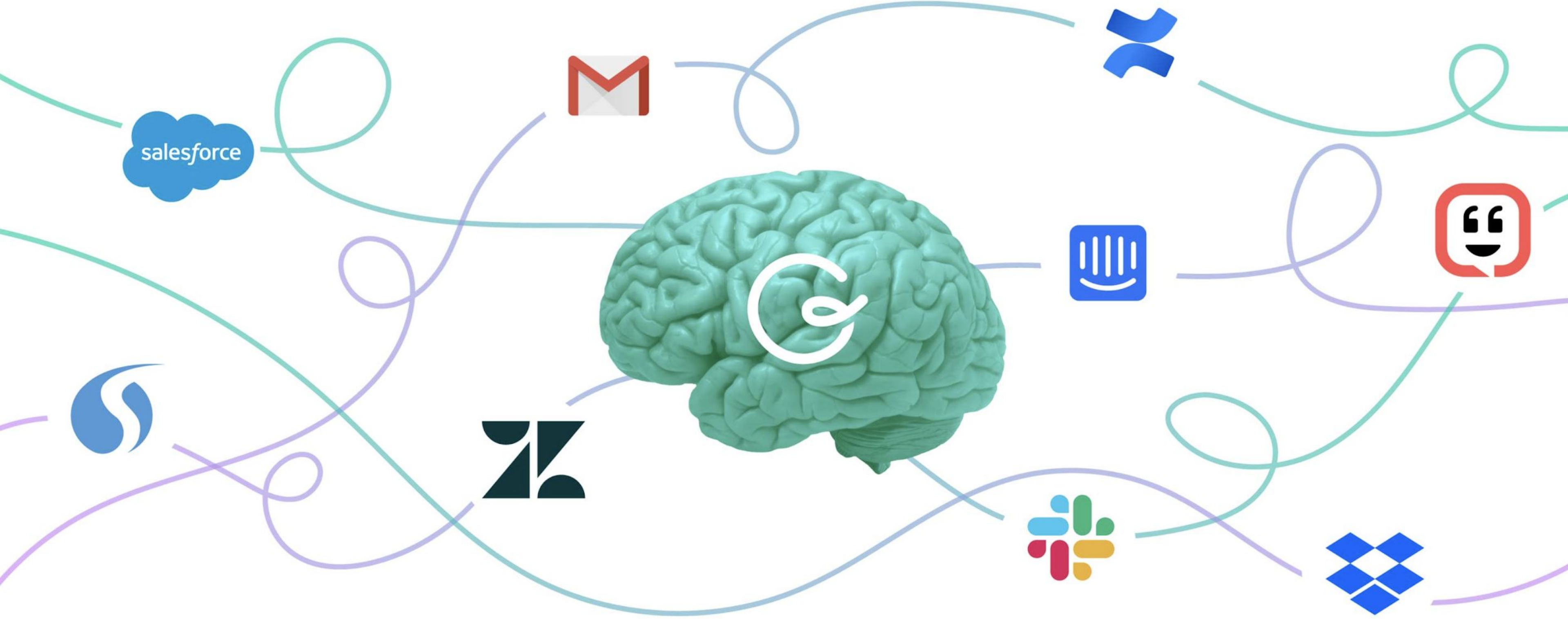
Yev Meyer
Sr. Data Scientist
Guru

GURU

Guru’s mission

GURU

We believe
the knowledge you need to do your job should find you


salesforce

Information workers switch windows on average **373 times per day** or around **every 40 seconds** while completing their tasks.

(Mark et al., 2016)
(Molla, 2019)

GURU

ML supporting the mission

Guru gathers **your company's knowledge** — from experts, documents, applications — and unifies it **into a single source of truth**.

Using ML, Guru then surfaces that knowledge to you in your favorite work applications (Slack, Intercom, Zendesk, Salesforce, Gmail, etc.)

GURU

A few ML features in production

AI Suggest Voice

suggest knowledge **in real-time** in phone conversations and conference calls

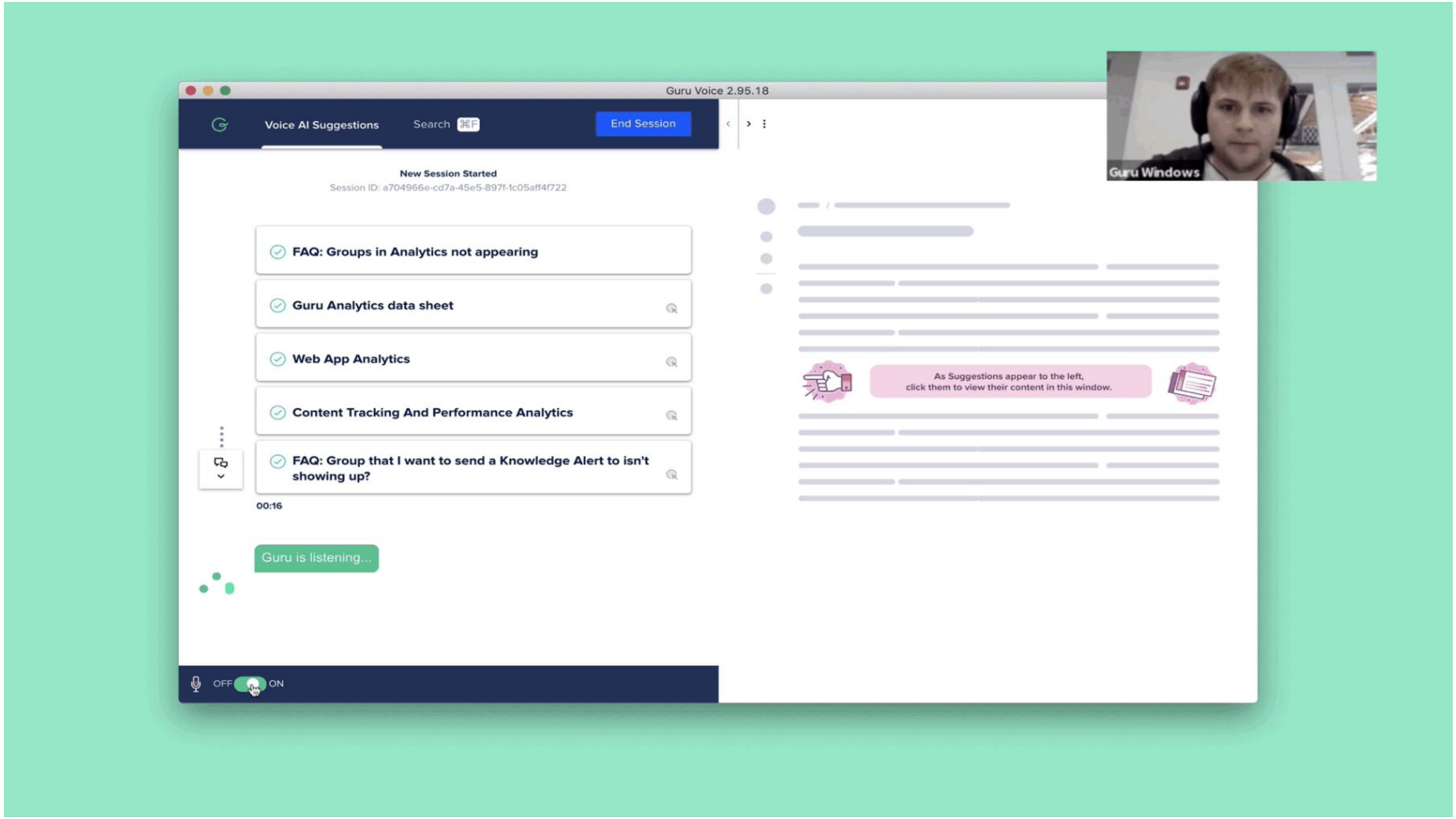
Listen to Audio → Transcribe Speech to Text → Recommend Knowledge

GURU

AI Suggest Voice

GURU

Demo


Guru Voice 2.95.18
Voice AI Suggestions
Search
End Session
New Session Started
Session ID: a704966e-cd7a-45e5-897f-1c05aff4f722
FAQ: Groups in Analytics not appearing
Guru Analytics data sheet
Web App Analytics
Content Tracking And Performance Analytics
FAQ: Group that I want to send a Knowledge Alert to isn't showing up?
00:16
Guru is listening...
OFF
ON
As Suggestions appear to the left, click them to view their content in this window.
Guru Windows

A hard problem to solve end-to-end

Client-side

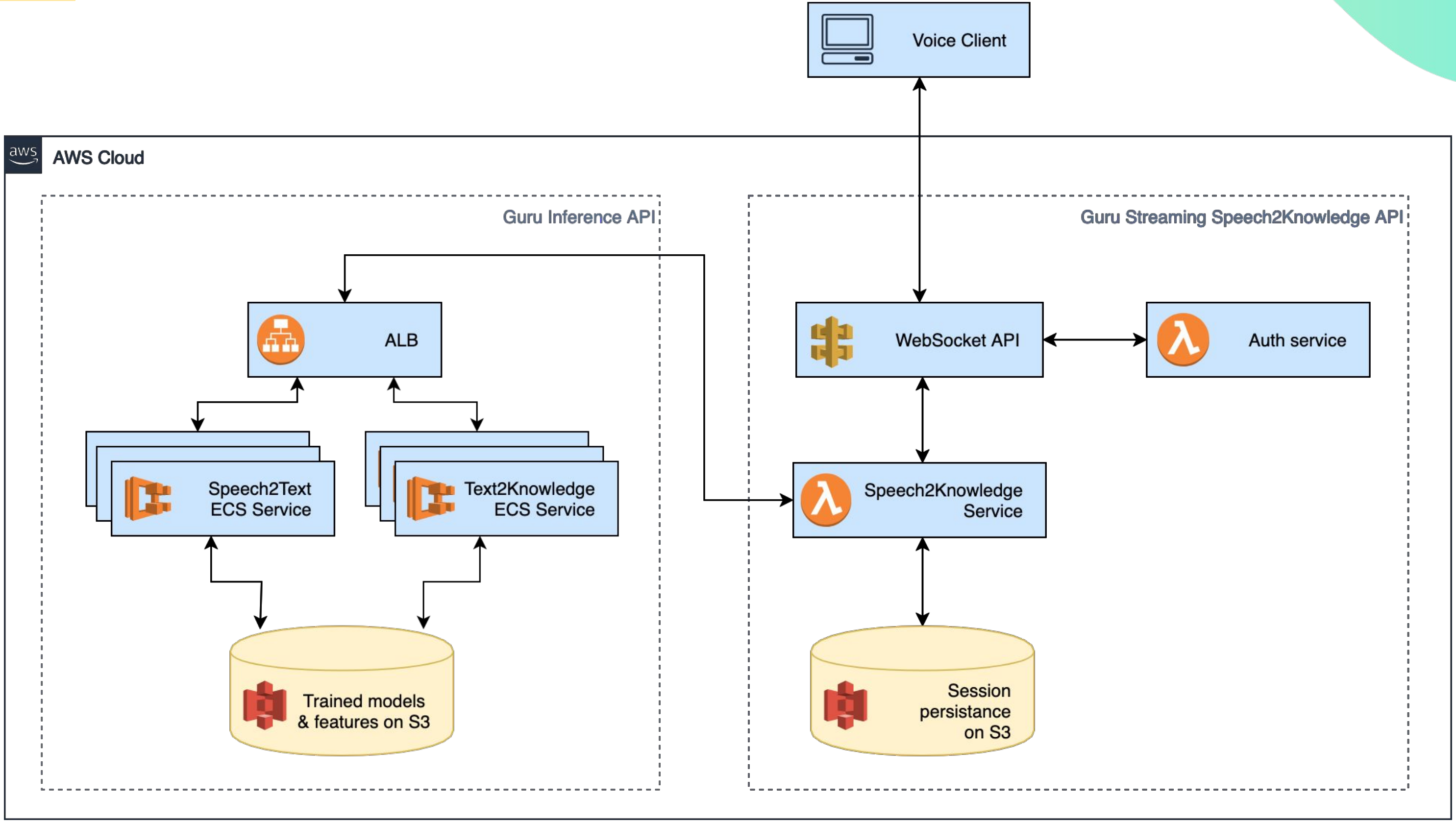
- capture audio for both parties (simplest case)
- stream all data in real-time
- support a variety of OS and hardware
- create UX that does not distract

DS-side

- transcribe speech and suggest knowledge, **all in real-time**
- handle speech detection, speaker separation, noise
- take custom jargon into account
- have scalable infrastructure for streaming, model training and serving
- embrace customer diversity: serve multiple models supporting the above
- make it cost-effective: GCP/AWS/Azure transcription is prohibitively expensive
    - added benefit: **specialized model**, built for a specific use-case
- get data for training the acoustic model

GURU

High-level architecture



Speech2Text service

Standing on the shoulders of giants. Literally.

Deep Speech: Scaling up end-to-end speech recognition

Awni Hannun*, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng

Baidu Research – Silicon Valley AI Lab

Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Vaino Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, Zhenyao Zhu

Baidu Silicon Valley AI Lab[1], 1195 Bordeaux Avenue, Sunnyvale CA 94086 USA
Baidu Speech Technology Group, No. 10 Xibeiwang East Street, Ke Ji Yuan, Haidian District, Beijing 100193 CHINA

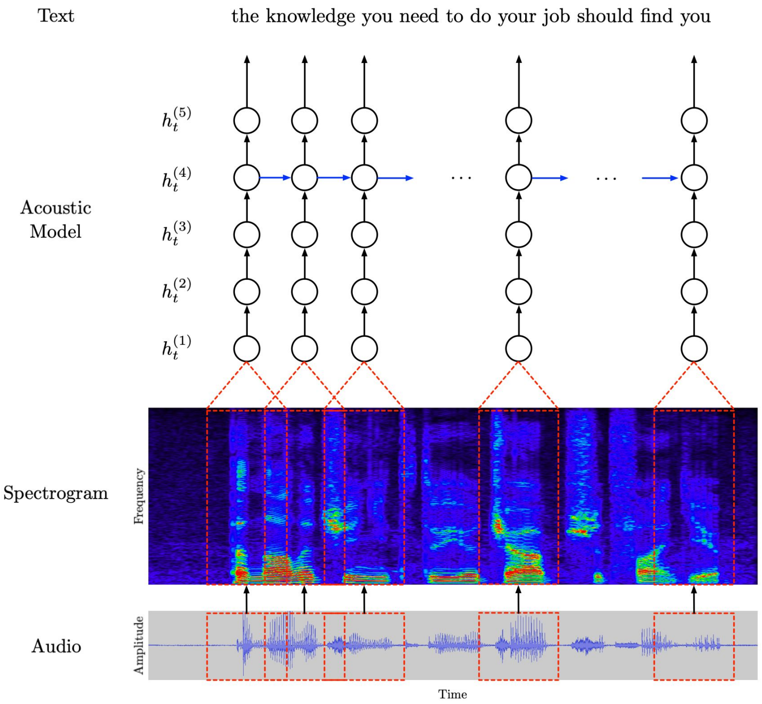
- Neural nets have been used in speech recognition for over 20 years
- However, there was no true **end-to-end** deep learning solution until ~2014
- Traditional systems employed heavily engineered processing stages, HMMs
- Baidu's was one of the first end-to-end demonstrations, predicting sequences of characters from input audio

⇒ Baidu's highly-simplified speech recognition pipeline has **democratized speech research**

⇒ Mozilla is one of the companies that was inspired to contribute to speech research

GURU

The approach: high-level



- Goal: given an utterance $x^{(i)}(t)$, $i = 1, \ldots, N$, generate a transcription sequence $\hat{y}^{(i)}$,

$$\hat{y}_\tau^{(i)} \in \{a, b, \ldots, z, ', \text{␣}\},\ \tau = 1, \ldots, T^{(i)}$$

- Approach: train a network that would allow us to extract $\hat{y}^{(i)}$ from the final layer
- Use RNN, with a **sequence of log-spectrograms**

$$x_{t,p}^{(i)}$$

as features, where *p* denotes the frequency band.

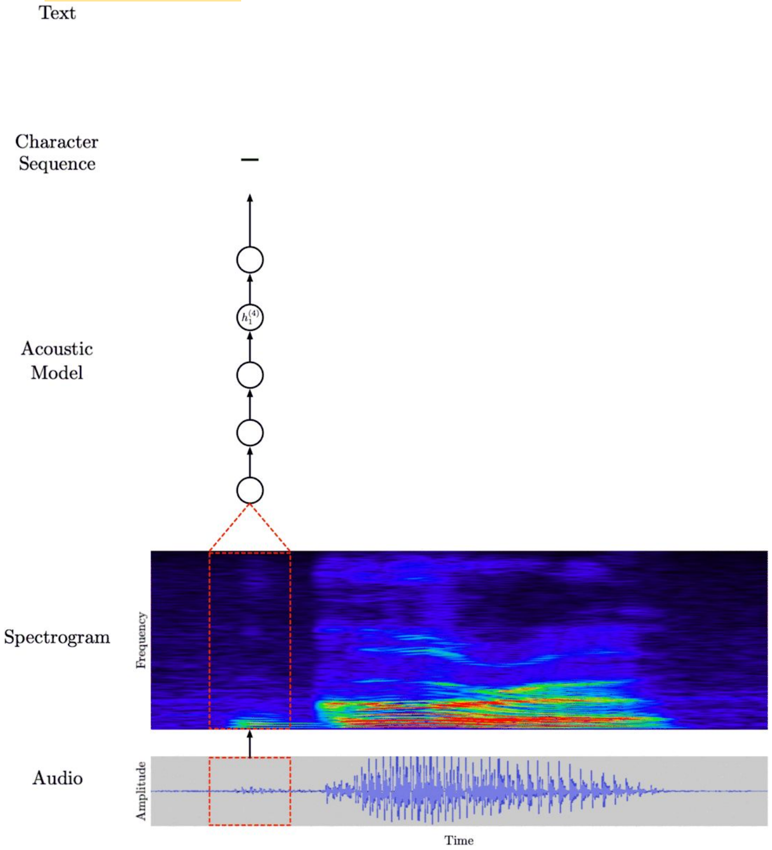
First three layers: non-recurrent, fully connected, taking neighboring context *C* into account

Fourth layer: **uni-directional recurrent**

Fifth layer: standard softmax

GURU

The approach: training



- The main challenge is that the transcription length stays the same across audio lengths
- We use connectionist temporal classification, or CTC (Graves et al., 2006)
- Layer 5 encodes a probability distribution $P(c \,|\, x)$ over **character sequences** c, where $len(c) = len(x)$

$$c_t \in \{a, b, \ldots, z, ', \text{␣}, \_\}$$

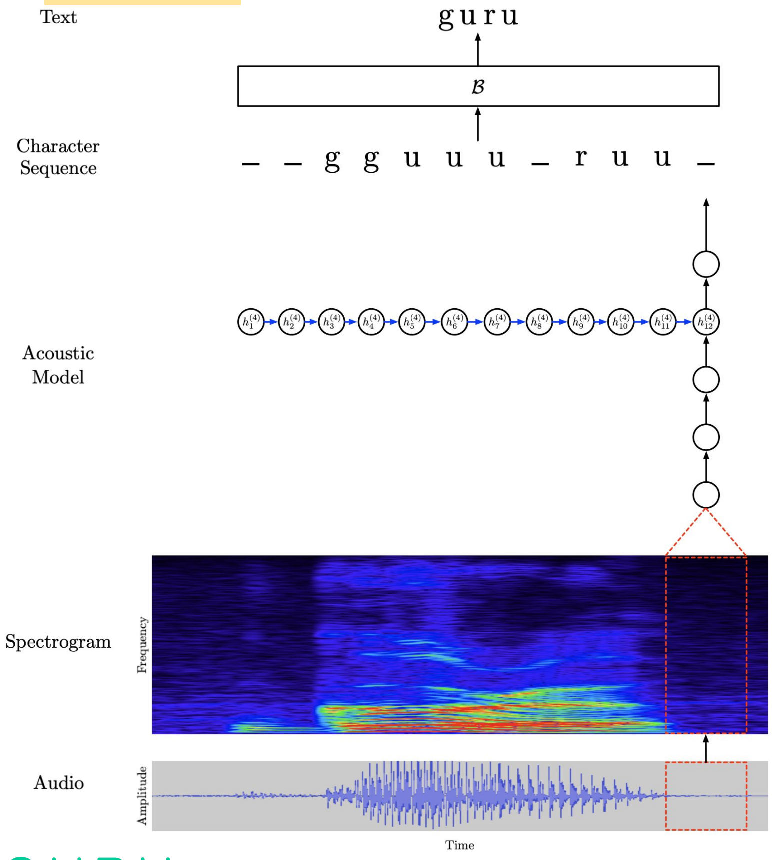
- Define a many-to-one map $\mathcal{B} : C \to Y$

$$y = \mathcal{B}(\text{“\_\_gguuu\_ruu\_”}) = \text{“guru”}$$

- Can now compute $P(y^{(i)} \,|\, x^{(i)}) = \sum_{c:\, \mathcal{B}(c) = y^{(i)}} P(c \,|\, x^{(i)})$
- Update parameters: $\theta^* = \arg\max_\theta P(y^{(i)} \,|\, x^{(i)})$

GURU

The approach: inference



- Decode the output, i.e., find the most likely transcription, e.g., by using max decoding via $\mathcal{B}$ or using prefix-decoding
- However, even with best decoding, you see spelling and linguistic errors (the "Tchaikovsky" problem)
  - Introduce a language model (LM)
  - We use an n-gram model (KenLM) that is trained on publicly available corpora
  - Can quickly look up words via beam search
  - Most importantly, can quickly update with new or newly-important words

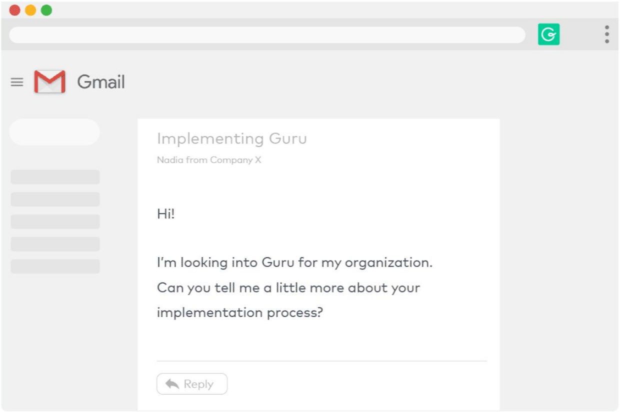
RNN output	Decoded Transcription
what is the weather like in bostin right now	what is the weather like in boston right now
prime miniter nerenr modi	prime minister narendra modi
arther n tickets for the game	are there any tickets for the game

GURU

Text2Knowledge service

GURU

Text2Knowledge



- **Offline:** run an NLP pipeline to extract features from individual pieces of knowledge (cards) and embed each card in a multi-dimensional space
- Use these features along with user-interaction data to train a weakly-supervised recommender system
- Weakly supervised, since not all interactions guarantee that a card was used in a conversation. In other words, the labels are noisy.
- **Online:** process newly-observed text using the same NLP pipeline and suggest top K cards.

Quick Recap

GURU

Quick Recap

- Our mission: **the knowledge you need to do your job should find you**
- AI Suggest Voice: applying the above to voice
- This is a hard problem to solve end-to-end
- Doable, given recent advances in e2e deep learning for speech recognition
- RNN + CTC + LM works really well
- Speech2Text + Text2Knowledge = Speech2Knowledge

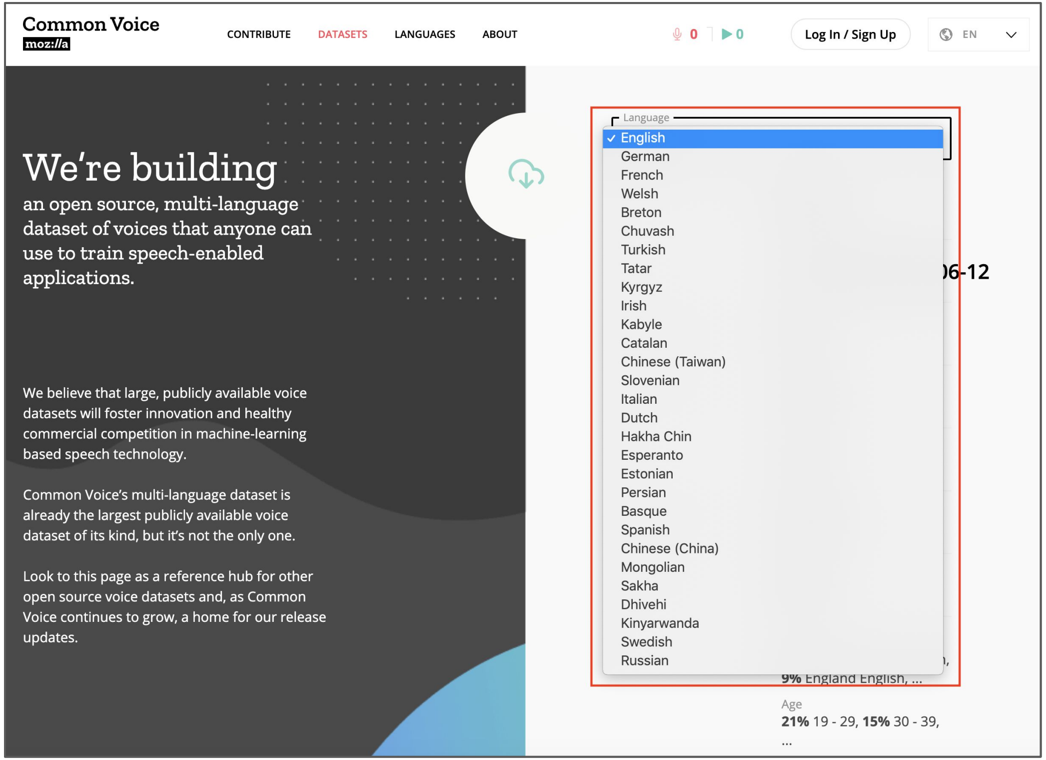
GURU

Lessons learned

GURU

Lessons learned: quality data is key

- The biggest challenge is having access to audio data for training
- Baidu's network was trained on more than **10k hours of audio**
- Mozilla realized that access to such data will allow for broad innovation in the space. Hence, **Common Voice**
- Can use other public data sets
- Can also synthesize data
- LM: quality data matters



Other lessons learned

- Audio packets coming from the client out of order
- Transcriptions being generated out of order
- Serverless VAD is a real challenge
- N-gram LMs are quite large
- Scalability lessons galore

- **Being gritty**
    - We are a small team, but we have grit

The most important slide

GURU

Everything discussed is a fruit of many people's labor at Guru.


Jenna Bellassai


Ed Brennan


Bernie Gray


Yev Meyer


Nabin Mulepati

Product Data Science Team

Come say hi and stop by our booth!

GURU

Thank you!

References

Mark G., Iqbal S., Czerwinski M., Johns P., Sano A. Neurotics Can't Focus: An in situ Study of Online Multitasking in the Workplace. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016.

Molla R. The productivity pit: how Slack is ruining work. Recode, 2019 https://www.vox.com/recode/2019/5/1/18511575/productivity-slack-google-microsoft-facebook. Accessed 12 Nov. 2019.

Hannun A., Case C., Casper J., Catanzaro B., Diamos G., Elsen E., Prenger R., Satheesh S., Sengupta S., Coates A., Ng A. Deep Speech: Scaling up end-to-end speech recognition. arXiv:1412.5567v2 [cs.CL], 2014.

Graves A., Fernández S., Gomez F., Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, ICML '06 Proceedings of the 23rd international conference on Machine learning