Speech to Text

The New interface for Communication

Joseph Fong

Background Information

- Startup Founder
- Systems Engineer
- Datacenter Infrastructure Manager
- Project Manager
- Technical Account Manager / customer Success Manager
- Product Marketing Manager
- Management Consulting – AI / ML startups
- Past VP STC San Francisco Chapter

Speech enabled services

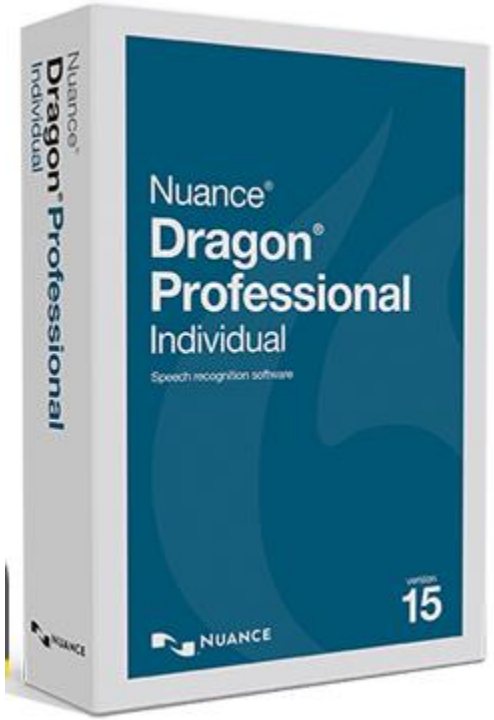
Microsoft
Cortana

Google Voice

Google Translate

Other Examples: Amazon Transcible API, Google Docs, MS Word, IBM Watson (free, 500 min audio input)

Software solutions for Speech to text



MaestraSuite

sonix

Capterra

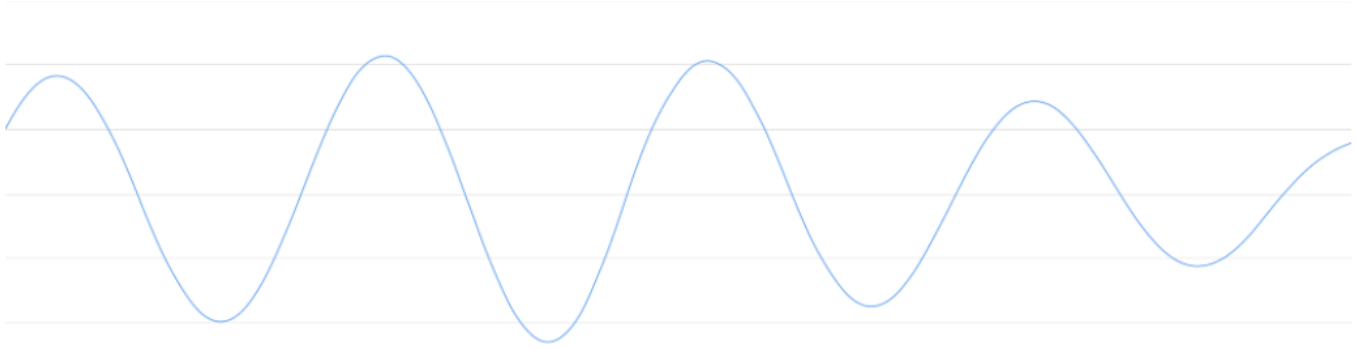
castel

Speech to Text market

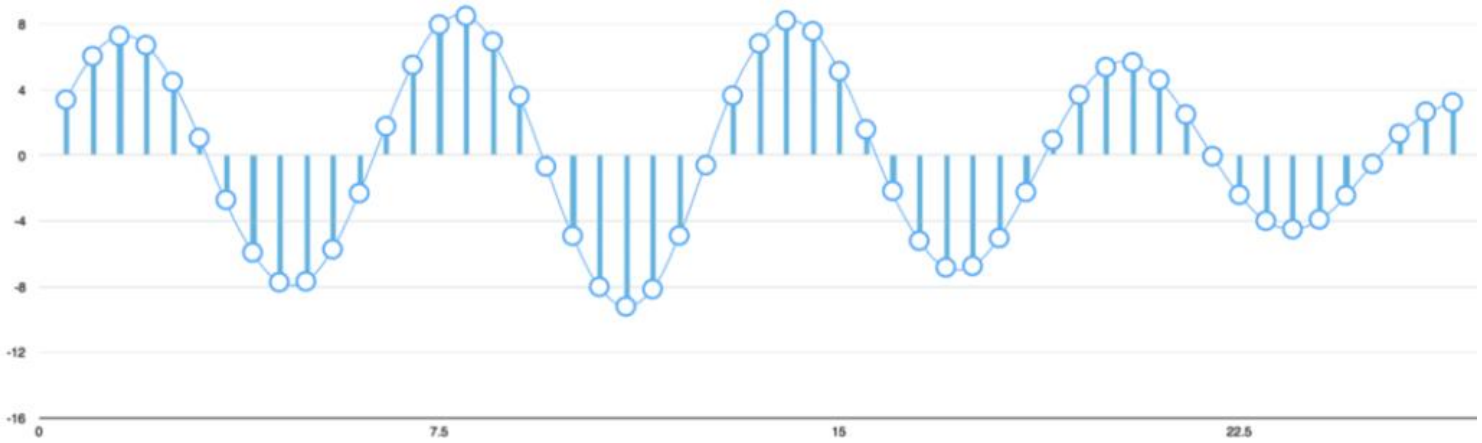
- Andrew Ng (Stanford) predicated that 99% reliability is required for critical mass
- Consumer market leading technology adoption
- Enterprise adoption will likely happen in the next 5-7 years
  - Lead by sectors with field operations
- National Privacy Laws protect service providers from liability

How Speech to Text Works

Converting sound waves into bits

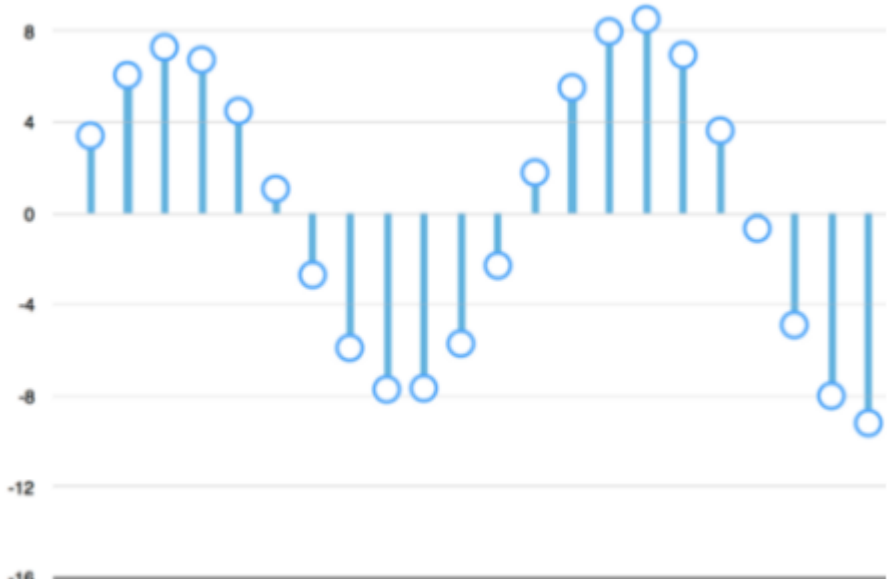


Sampling



Sampling Rate

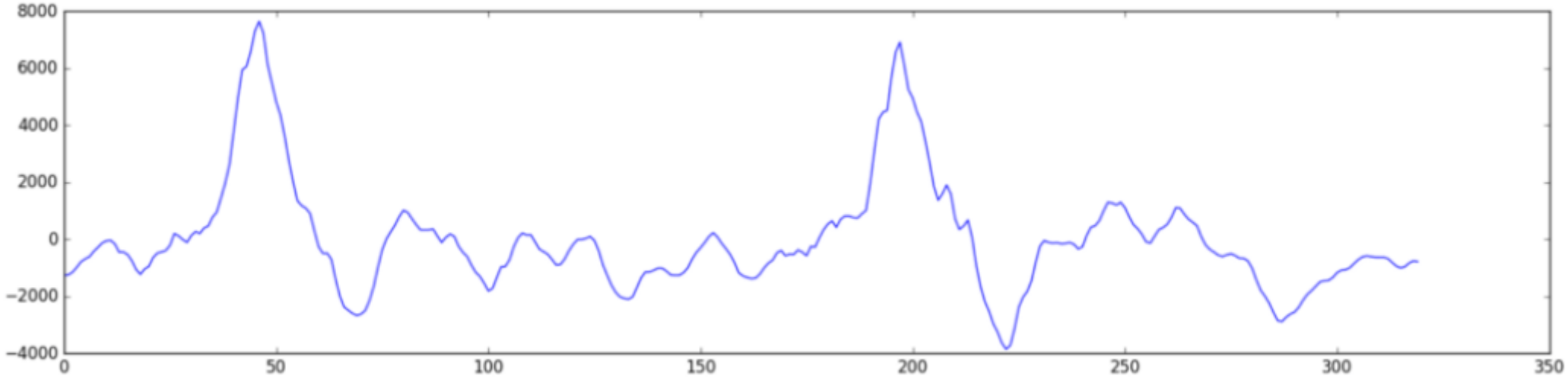
- CD quality = 44khz (44,000 samples per sec)
- Human voice = 16khz (16,000 samples per sec)



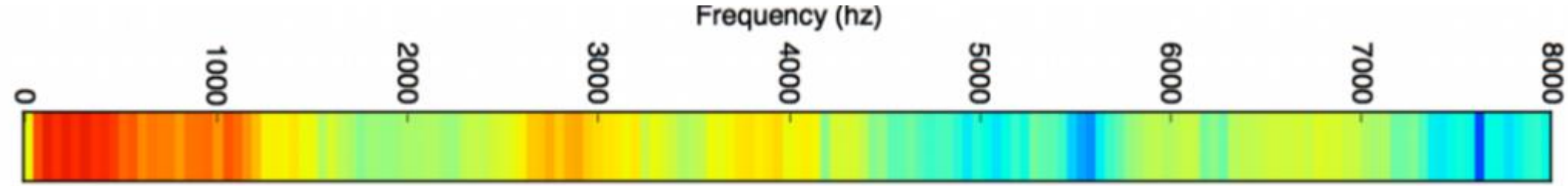

```
[-1274, -1252, -1160, -986, -792, -692, -614, -429, -286, -134, -57, -41, -169, -456, -450, -541, -761, -1067, -1231, -1047, -952, -645, -489, -448
, -397, -212, 193, 114, -17, -110, 128, 261, 198, 390, 461, 772, 948, 1451, 1974, 2624, 3793, 4968, 5939, 6057, 6581, 7302, 7640, 7223, 6119, 5461,
 4820, 4353, 3611, 2740, 2004, 1349, 1178, 1085, 901, 301, -262, -499, -488, -707, -1406, -1997, -2377, -2494, -2605, -2675, -2627, -2500, -2148, -
1648, -970, -364, 13, 260, 494, 788, 1011, 938, 717, 507, 323, 324, 325, 350, 103, -113, 64, 176, 93, -249, -461, -606, -909, -1159, -1307, -1544]
```

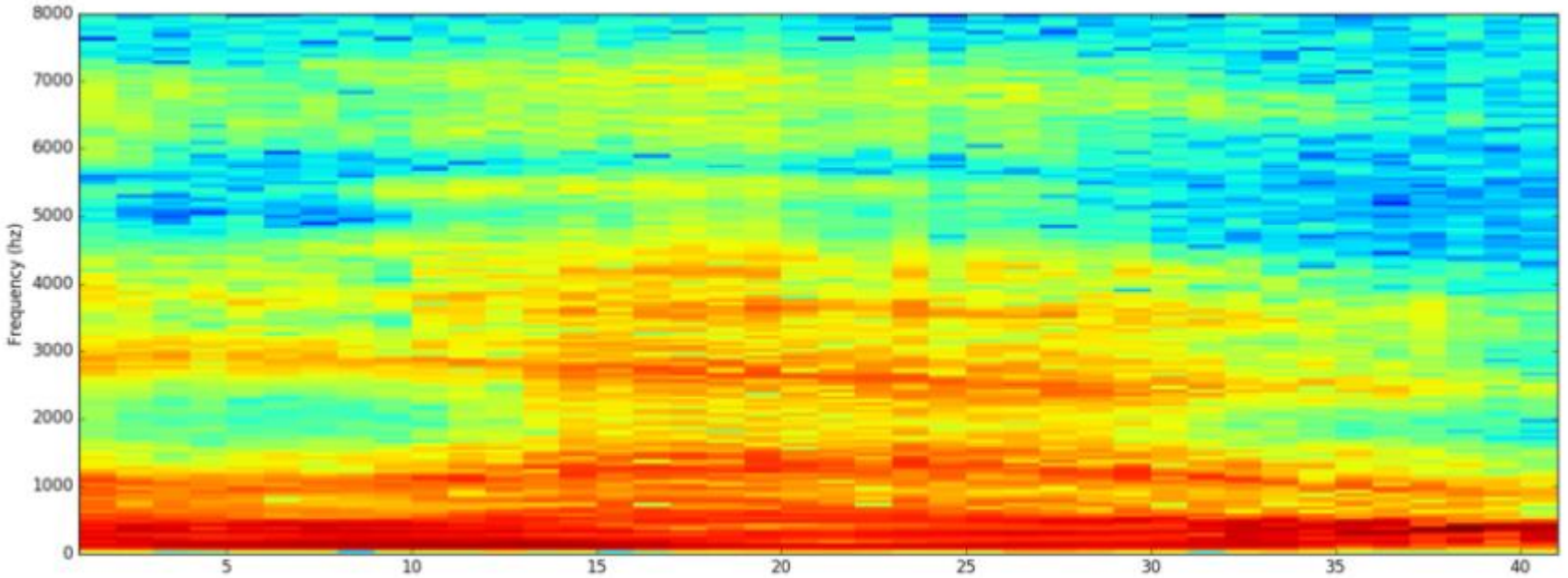
Each number represents the amplitude of the sound wave at 1/16000th of a second intervals

Preprocessing

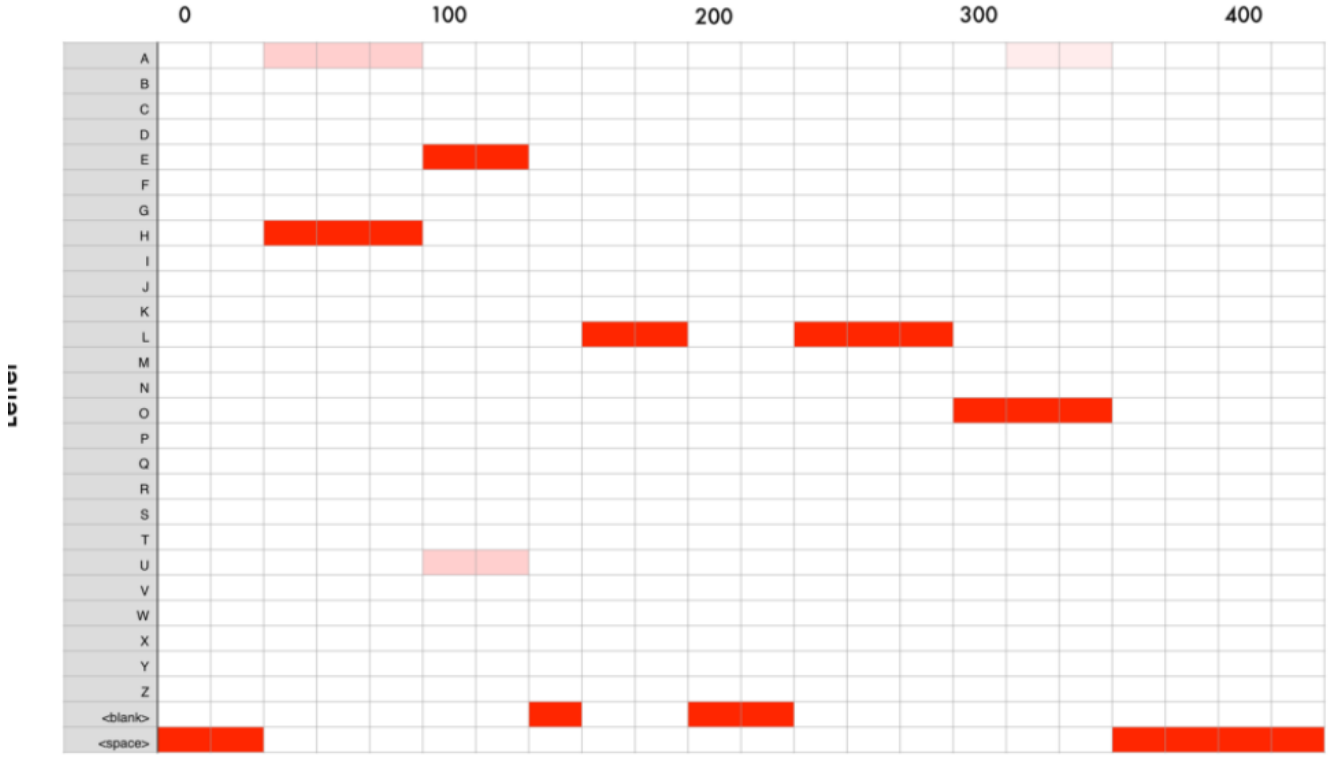


Analysis using Heat Index





Analysis over time (the word Hello)



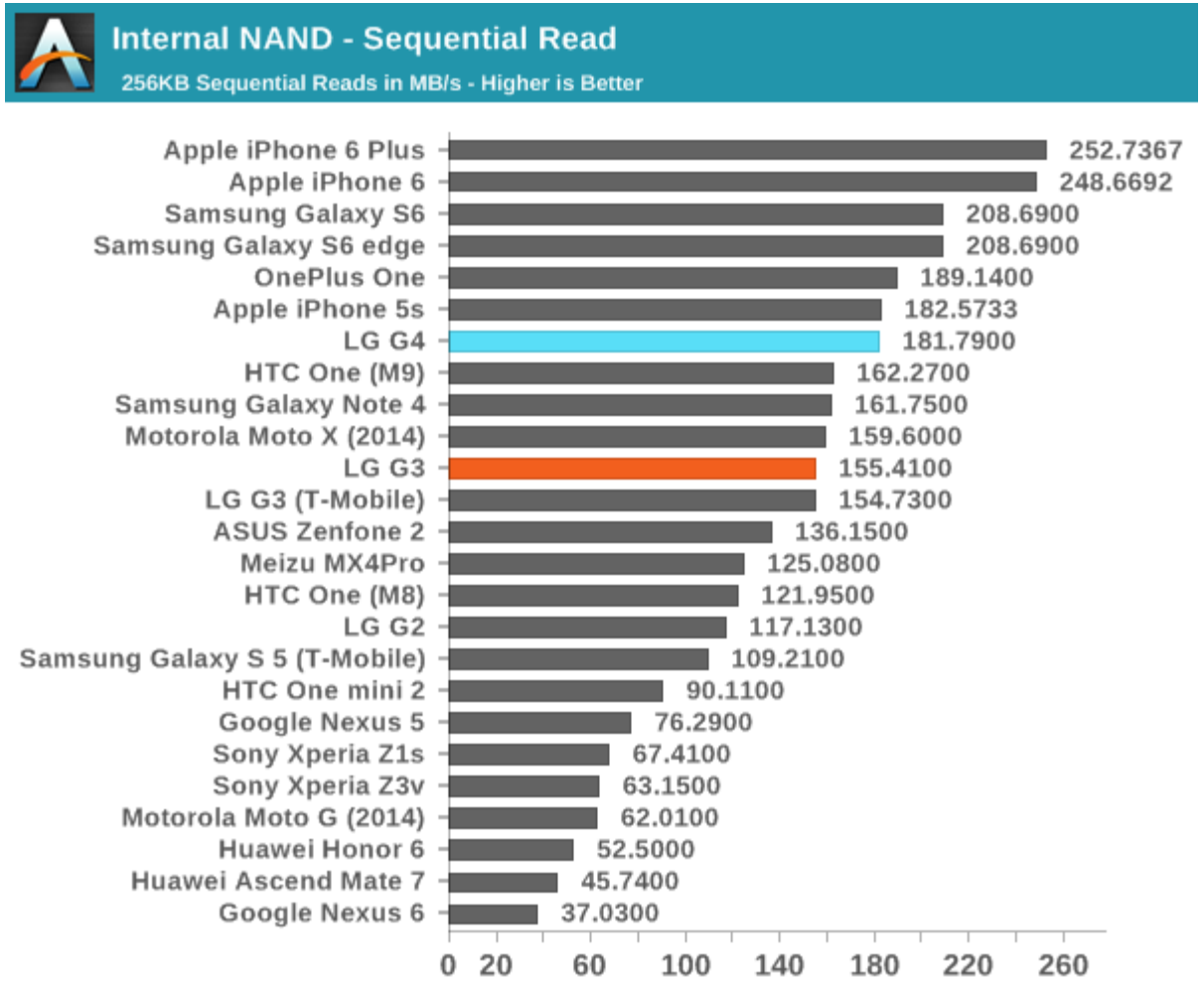
Speech is sample using an interval: ex. 20ms
Through analysis of literary data, the computer predicts that HE_L_LO is actually **HELLO**

Speech to text – game of probabilities

- Understanding Contextual speech is very difficult
- Someone may say Hullo instead of Hello
- The computer may predict you said BYE when you said BY
- Time needed to adjust to speaker’s speaking style, voice, etc
- Emotions and Accents interfere speech recognition
- Speaking rate – words per sec
- Localization is challenging

Compute resources

Mobile phones – wide range of compute capabilities



Real time processing challenges

Device based processing is advantageous

- Siri processes speech locally
- Google Home is primarily cloud based (high latency)
  - Internet connection required

Background Noise (multiple speakers, ambient noise)

- filtering reduces error rates
- Trade off is reduced real-time processing accuracy

Machine Learning – secret sauce

- ML is required to deliver complex speech to text processing
- Service being provided by startups and Enterprises
- Advantageous to subscribe to API speech services
  - Ex. IBM Watson, AWS, Knowles
- API usage allows for rapid integration into your environment
- API training and document is readily available
- Higher accuracy than what you build yourself
- More cost effective and faster time to market

Privacy considerations to using speech tools

- Recordings must comply with state laws
- California – all parties must agree
- Texas – only one party needs to agree
- Presumption of privacy – public vs private spaces
- Conference calls that are multi-regional
- GDPR

Thank you

Discussion / QA

sflife1@Hotmail.com
www.linkedin.com/in/jofong