Voice Recognition

By

Luigi Penaloza

Voice recognition, automatic/computer speech recognition is the process of converting acoustic signal captured by microphone into set of words.

Voice recognition:

- Recognize the voice of the individual
- Identifies what you are saying
- Performance of required task by computer

Voice Recognition History

Humans vs Computers

When we articulate, sound waves are produced and ear conveys to brain for processing.

- Digitization
- Acoustic Analysis of speech signal
- Linguistic interpretation

1940’s and 1950’s- US Department of Defense and MIT largest contributors to the voice recognition field.

1960’s and 1970’s- Technology attributed by educational institutions.

Voice Recognition now

- Built into our phones, games consoles and smart watches.
- Used to automate our house.
- Integrated into personal AI assistants, that allows us to shop online and make appointments.

All possible because deep learning has made voice recognition accurate enough.

Right now we have about 95% word accuracy for the English language.

Famous Voice Assistants

- Siri (Apple)
- Duer (Baidu)
- Cortana (Microsoft)
- Alexa (Amazon)
- Google Now (Google)
- Watson (IBM)

amazon alexa

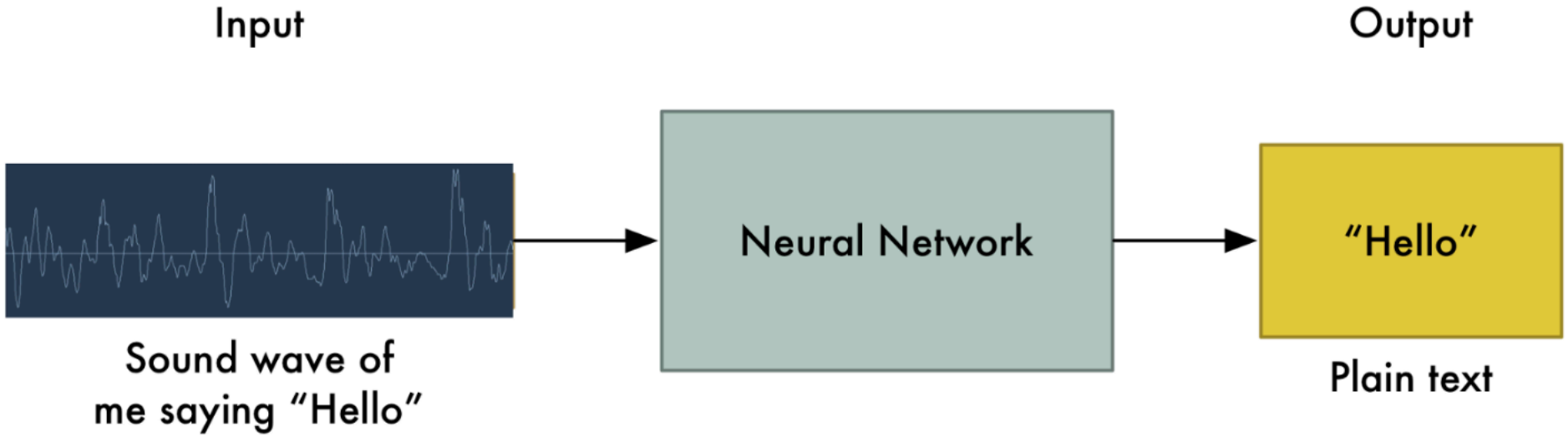
IBM WATSON

Cortana

Hey Siri

How does it work?

- The user speaks some words by invoking voice recognition on a mobile app.
- The spoken words are processed by the recognition software and converted to text.
- The converted text is then provided as input to the search mechanism, which returns the results.

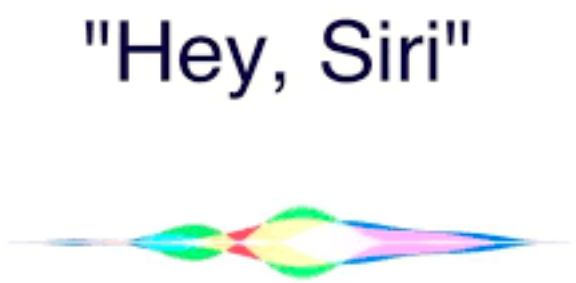


Difficulties with speech recognition

Main problem is that speech varies in speed.

For example: one person might say “hello Siri!”, and another might say ”heeeellllllooooo Siiiiriii!”

Both files should be recognized as same text, “hello Siri!” so we need to align audio files of different lengths to a fixed length piece.



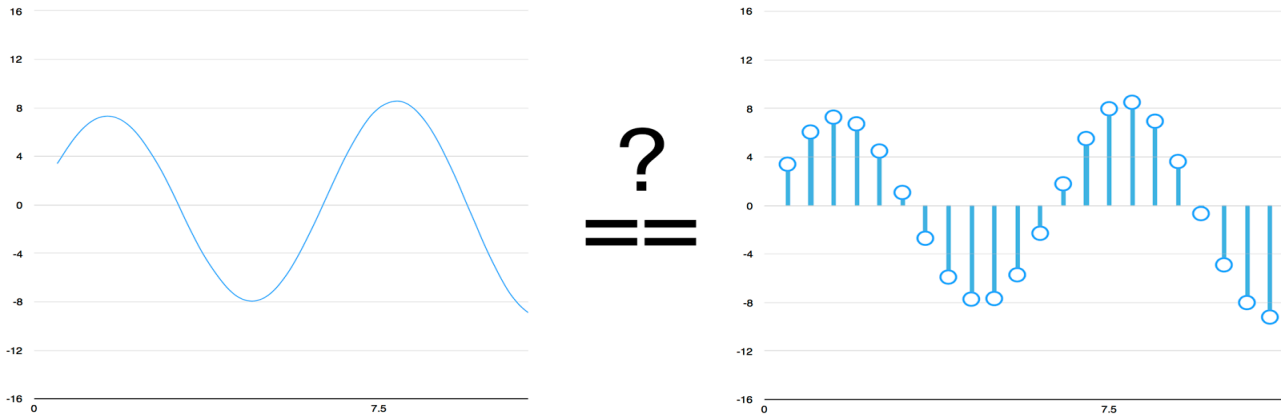
Turning sound into bits

Sampling

- We feed sound waves into a computer. Since sound waves are one dimensional, we can obtain a single value based on height of wave.

- We record the height of wave at equally spaced points.
- We can take a reading a thousand times of a second and record wavelength height number.

- 16,000 samples per second is enough to cover the frequency range of human speech.



Original Analog Signal

Sampled Digital Signal

[-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]

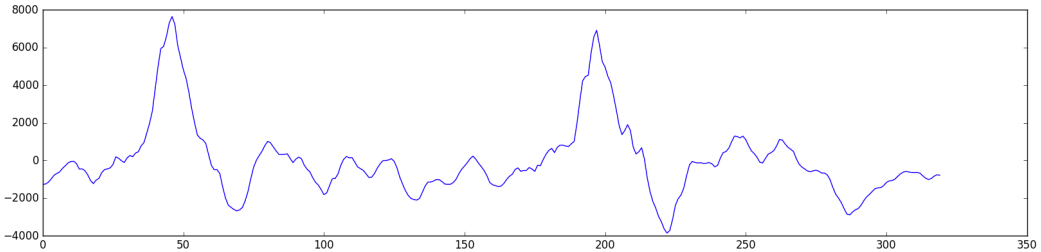
Each number represents the amplitude of the sound wave at 1/16000th of a second intervals

Pre-processing sampled data

- We have an array of number representing the sound wave's amplitude.

in our case 1/16,000th of second intervals.

- We can preprocess the data and then feed numbers to neural network.



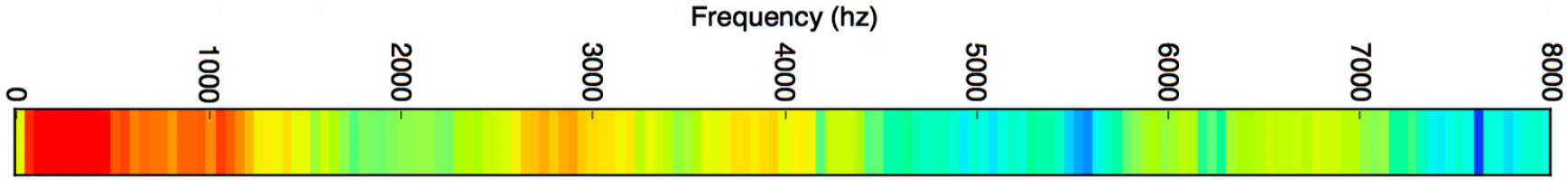
[-1274, -1252, -1160, -986, -792, -692, -614, -429, -286, -134, -57, -41, -169, -456, -450, -541, -761, -1067, -1231, -1047, -952, -645, -489, -448
, -397, -212, 193, 114, -17, -110, 128, 261, 198, 390, 461, 772, 948, 1451, 1974, 2624, 3793, 4968, 5939, 6057, 6581, 7302, 7640, 7223, 6119, 5461,
4820, 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,
-1815, -1725, -1341, -971, -959, -723, -261, 51, 210, 142, 152, -92, -345, -439, -529, -710, -907, -887, -693, -403, -180, -14, -12, 29, 89, -47, -
398, -896, -1262, -1610, -1862, -2021, -2077, -2105, -2023, -1697, -1360, -1150, -1148, -1091, -1013, -1018, -1126, -1255, -1270, -1266, -1174, -10
03, -707, -468, -300, -116, 92, 224, 72, -150, -336, -541, -820, -1178, -1289, -1345, -1385, -1365, -1223, -1004, -839, -734, -481, -396, -580, -52
7, -531, -376, -458, -581, -254, -277, 50, 331, 531, 641, 416, 697, 810, 812, 759, 739, 888, 1008, 1977, 3145, 4219, 4454, 4521, 5691, 6563, 6909,
6117, 5244, 4951, 4462, 4124, 3435, 2671, 1847, 1370, 1591, 1900, 1586, 713, 341, 462, 673, 60, -938, -1664, -2185, -2527, -2967, -3253, -3636, -38
59, -3723, -3134, -2380, -2032, -1831, -1457, -804, -241, -51, -113, -136, -122, -158, -147, -114, -181, -338, -266, 131, 418, 471, 651, 994, 1295,
1267, 1197, 1291, 1110, 793, 514, 370, 174, -90, -139, 104, 334, 407, 524, 771, 1106, 1087, 878, 703, 591, 471, 91, -199, -357, -454, -561, -605,
-552, -512, -575, -669, -672, -763, -1022, -1435, -1791, -1999, -2242, -2563, -2853, -2893, -2740, -2625, -2556, -2385, -2138, -1936, -1803, -1649,
-1495, -1460, -1446, -1345, -1177, -1088, -1072, -1003, -856, -719, -621, -585, -613, -634, -638, -636, -683, -819, -946, -1012, -964, -836, -762,
-788]

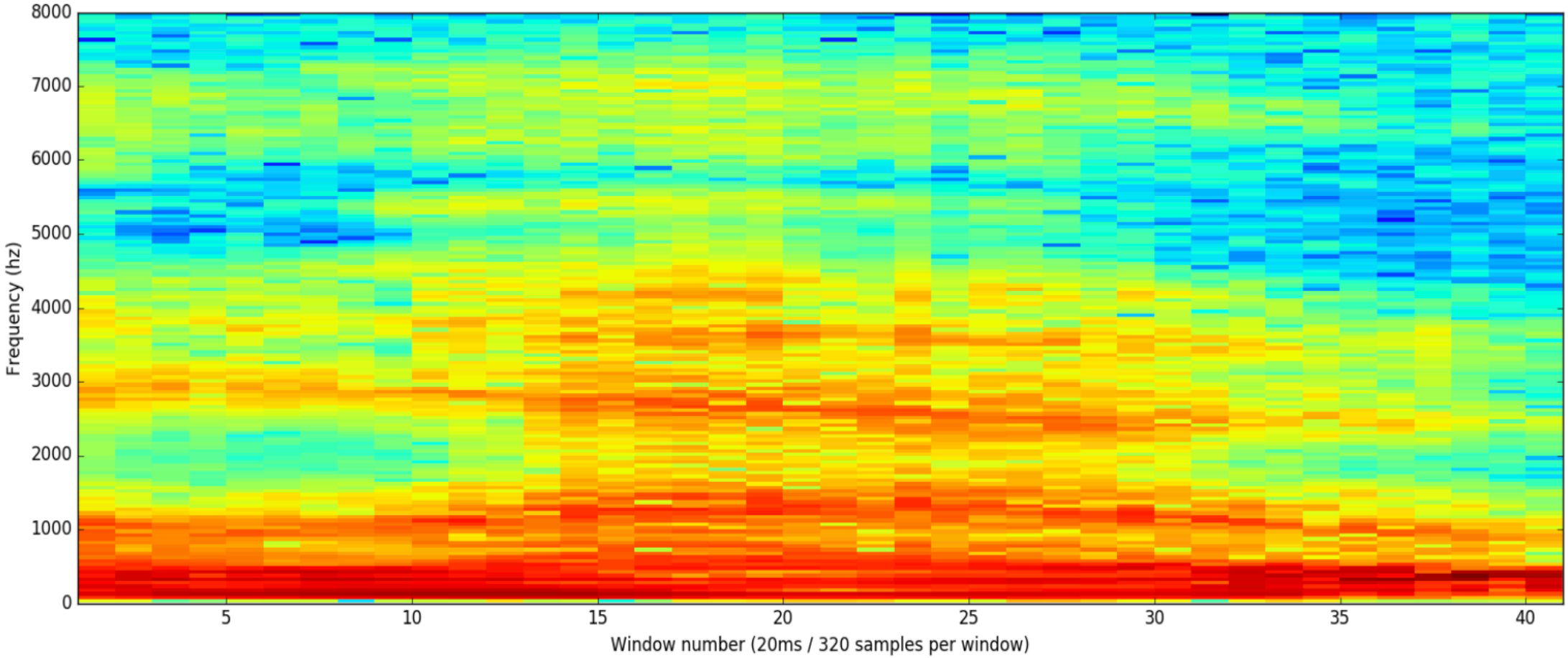
- First we group sampled audio into chunks
- Recordings usually mix of different frequen-

ices, low and high pitched.

- We can use Fourier Transform which

breaks apart complex sound waves into simple sound waves. Then we add how much energy is contained in each wave.

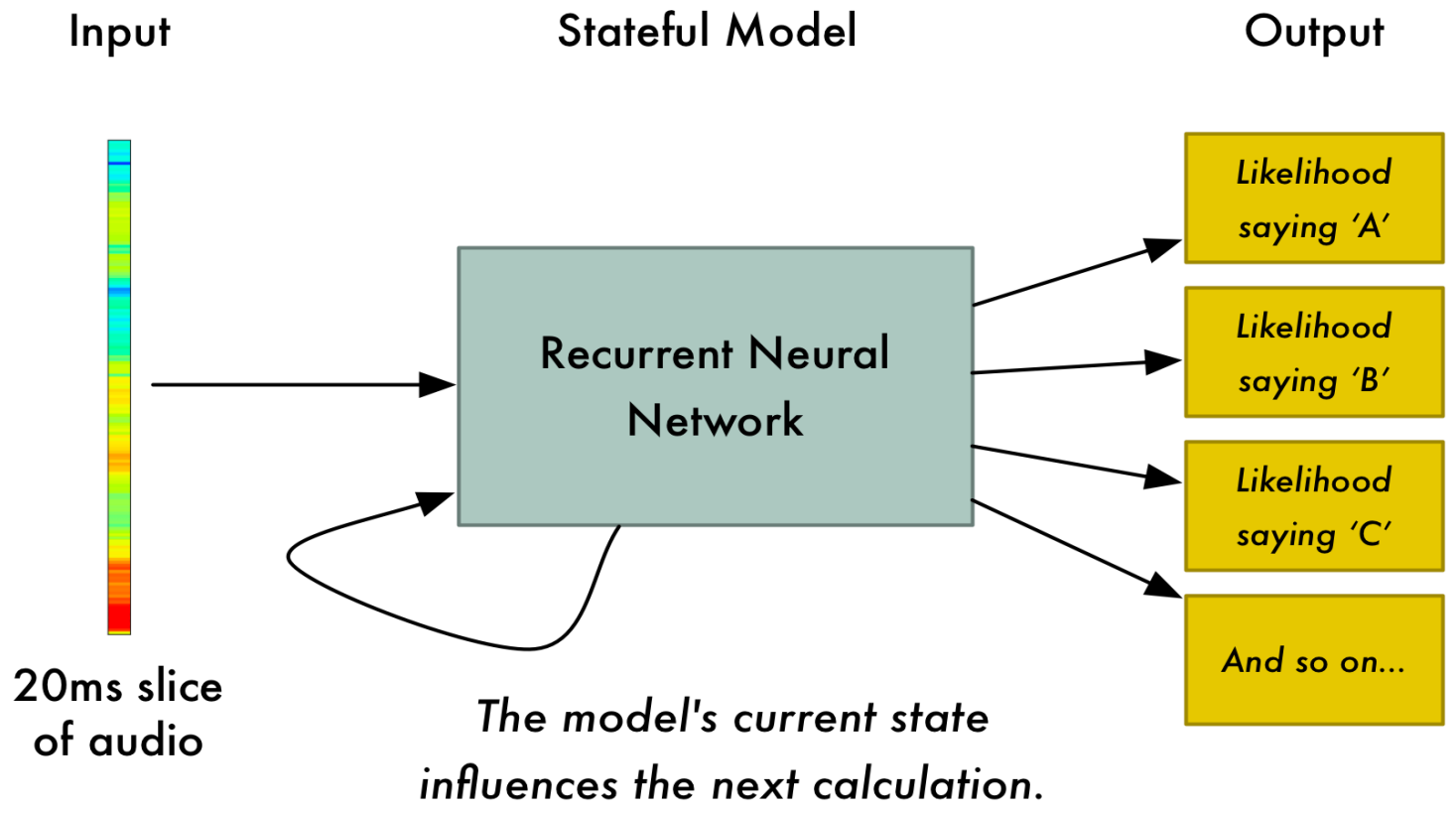


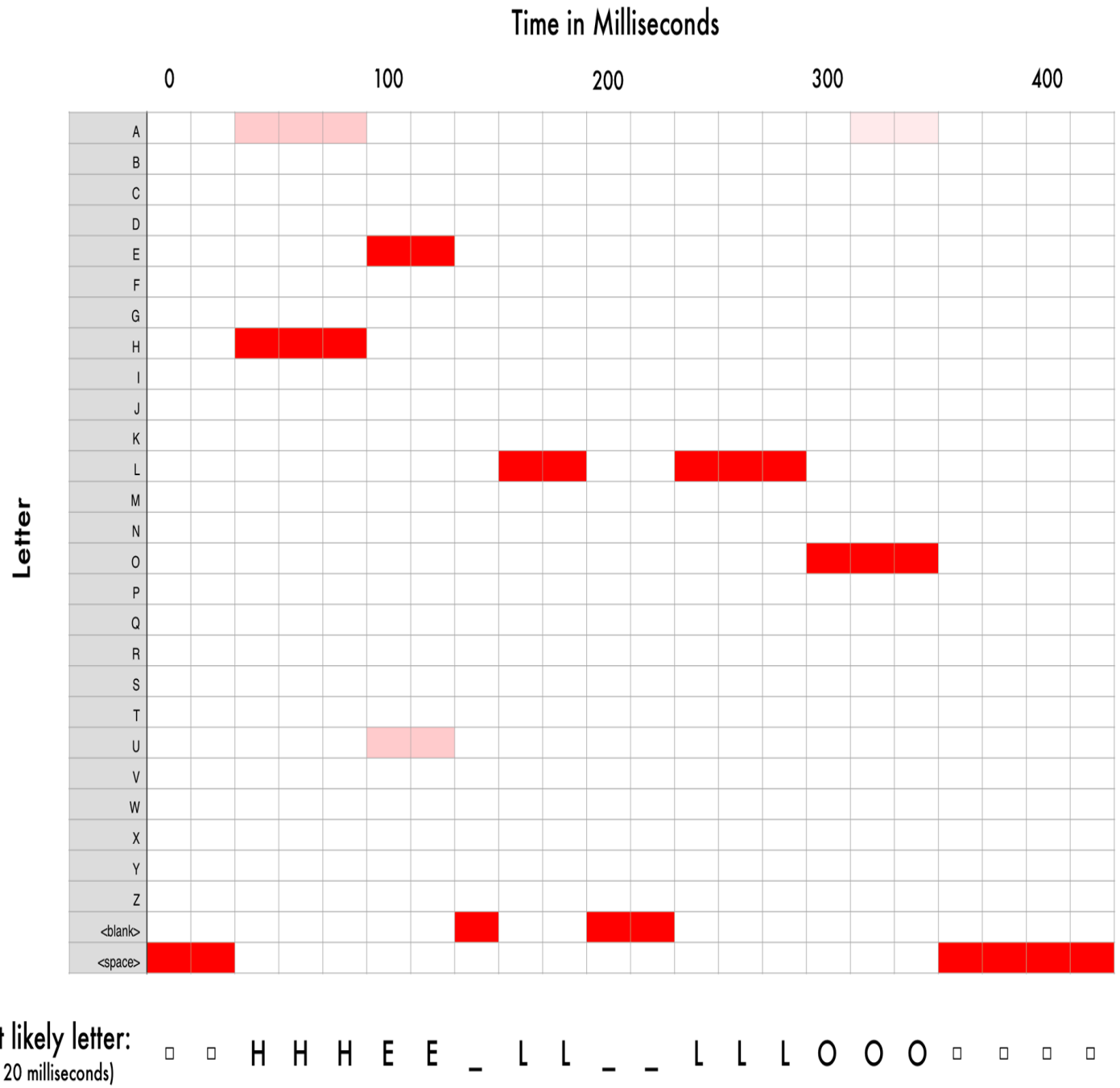


The full spectrogram of the "hello" sound clip

Recognizing characters from short sounds

- We'll feed audio in previous format to neural network
- Neural network input = 20 millisecond audio chunks
- For each audio, model will try to figure out the letter that corresponds to sound being spoken.





- Neural network thinks we said
- “HHHEE_LL_LLLOOO” or “AAAUU_LL_LLLOOO”
- With cleaning:
- “HHHEE_LL_LLLOOO”-> ”HE_L_LO”
- “AAAUU_LL_LLLOOO”->

“AU_L_LO”

“HELLO” or ”AULLO”

”HELLO”

Future of Voice Recognition

Competitiveness between voice assistants, and mobile app development companies will drive voice recognition advancement ahead.

SOURCES:

http://research.baidu.com

http://web.mit.edu

https://medium.com

www.sas.upenn.edu

www.voicerecognition.com.au

Luigi Penaloza