SAY WHAT?

Accent Classification for Native and Non-Native English Speakers

MSDS2020 Machine Learning Project

The Big 3 Tech Companies

…are heavily invested in voice recognition

Google Assistant

amazon alexa

Google Assistant

“OK Google, are there any restaurants near me?”

Google Assistant

“OK Google, can I run multiple n_jobs parameters within sklearn Grid Search to make my model run faster???

Don’t.”

Voice is the Future

…why type when you can talk?

TECHNOLOGY

3 Ways the Voice Revolution Is Going to Change Your Life

Be prepared for voice assistants to be everywhere in the future.

By Ken Sterling *Executive vice president, BigSpeak* @ken_sterling

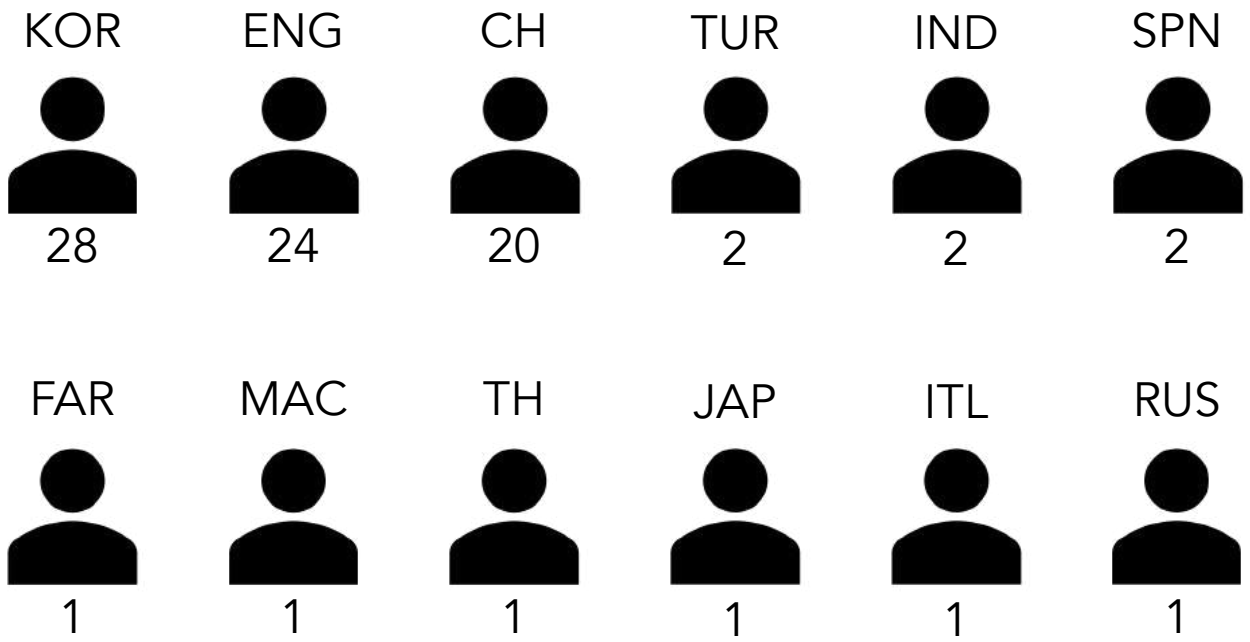
Accent detection can help fine tune recommendations and improve accuracy of speech predictions

Accent is a better indicator of cultural background

The Dataset

Wildcat Corpus of Native- and Foreign-Accented English

84 Participants

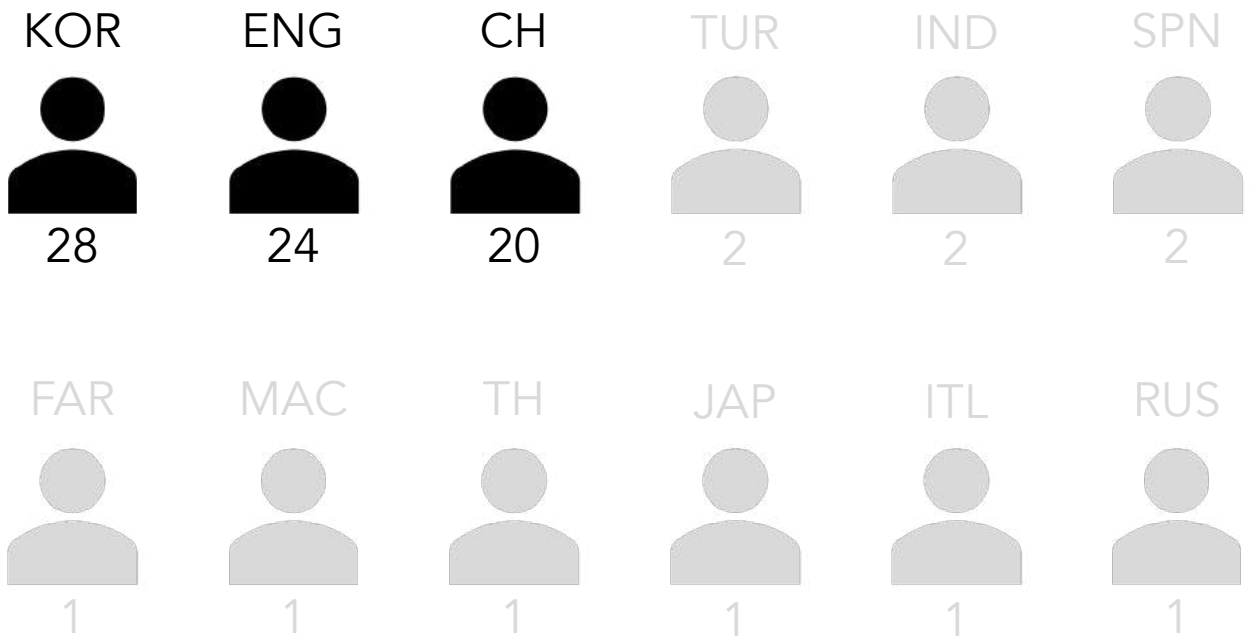


“Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.”

The Dataset

Wildcat Corpus of Native- and Foreign-Accented English

84 Participants

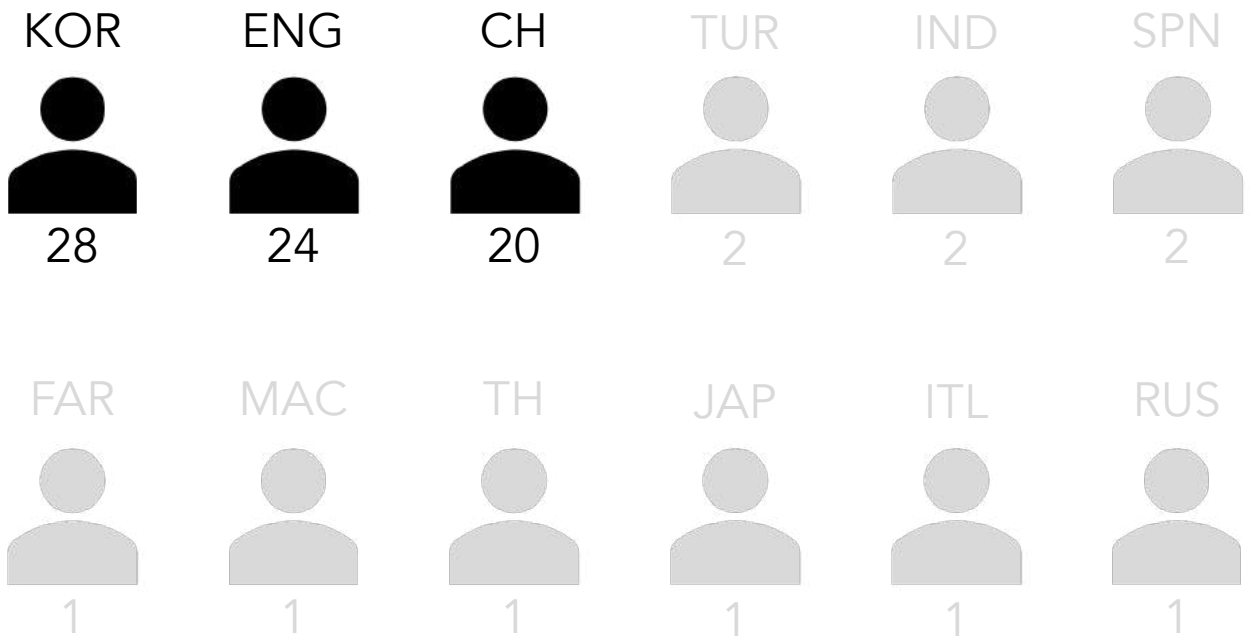


“Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.”

The Dataset

Wildcat Corpus of Native- and Foreign-Accented English

84 Participants



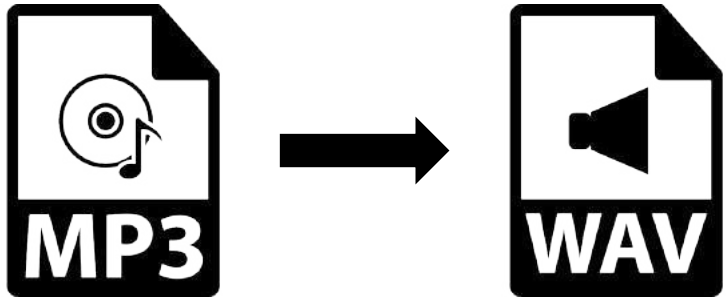
72

Data Points

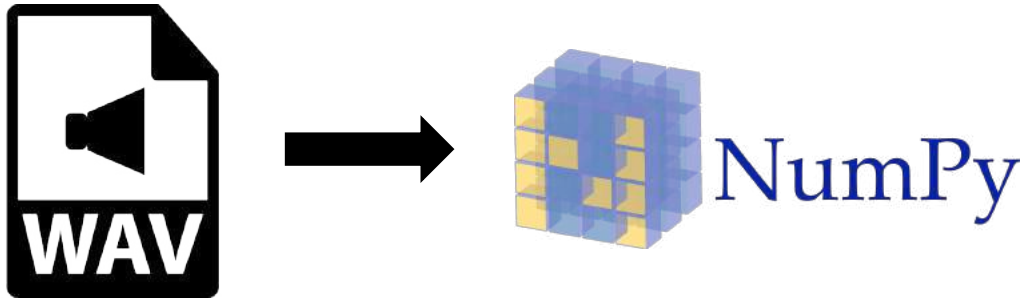
…we'll get to that.

Uhm Kyle, how do you process audio?

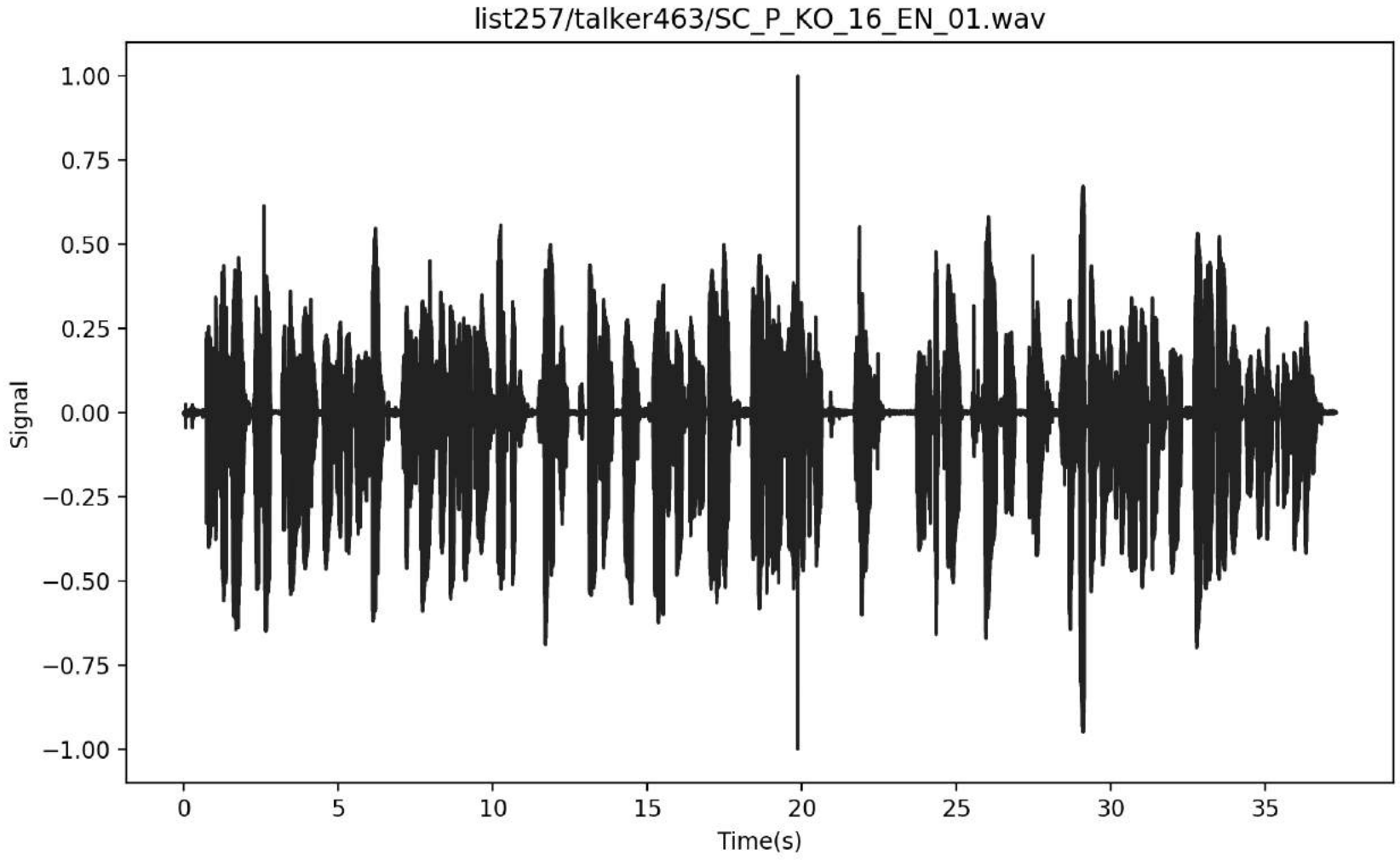
```
from pydub
import AudioSegment
```



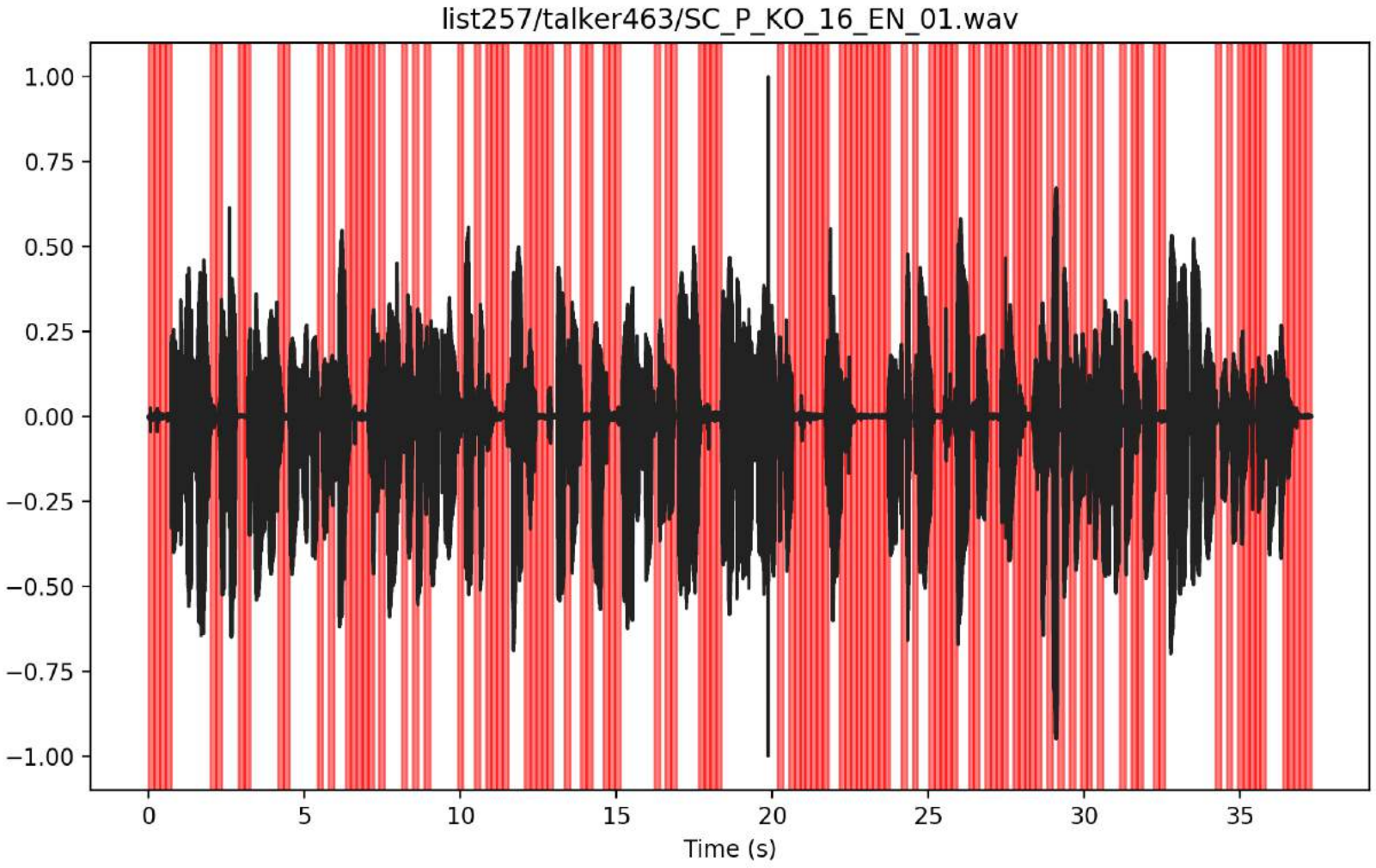

```
from librosa
import librosa.display
```



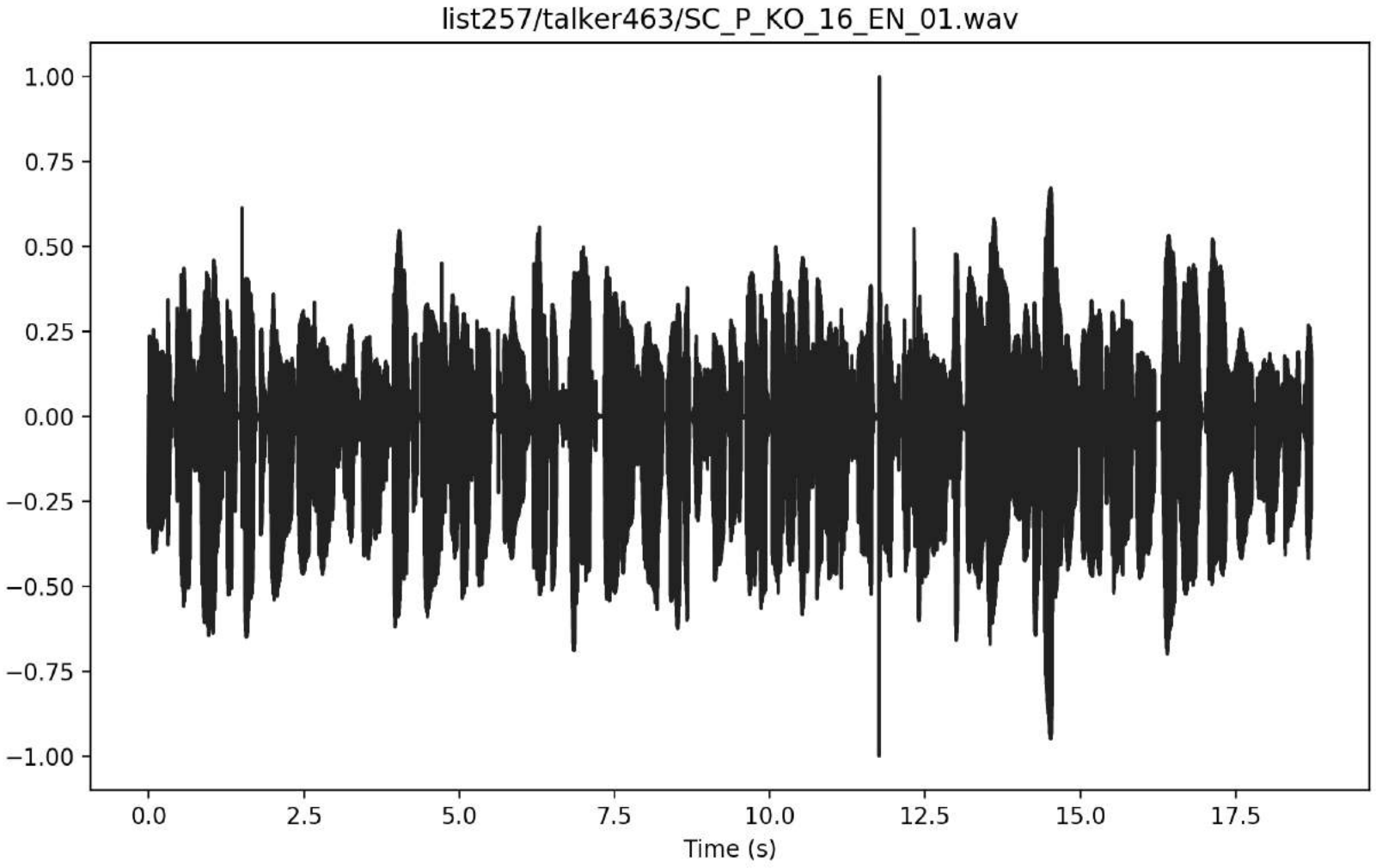
A peek into one file…

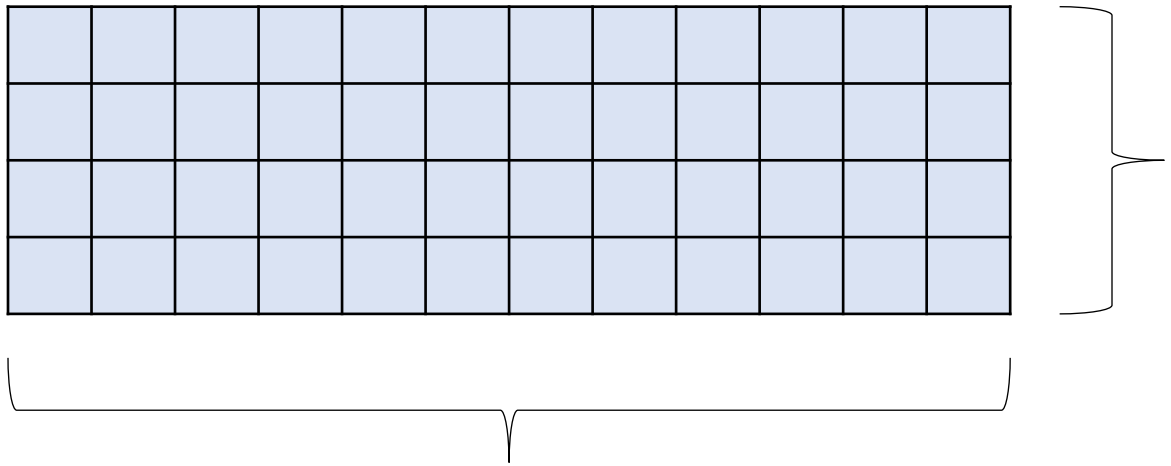


Let’s get rid of those!

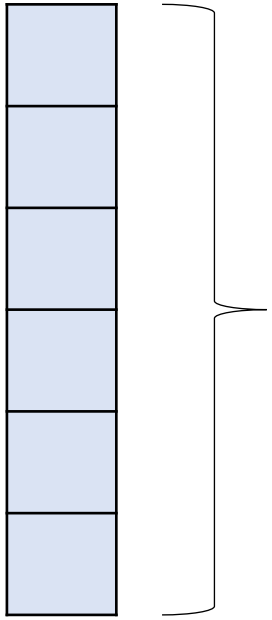


Let’s get rid of those!



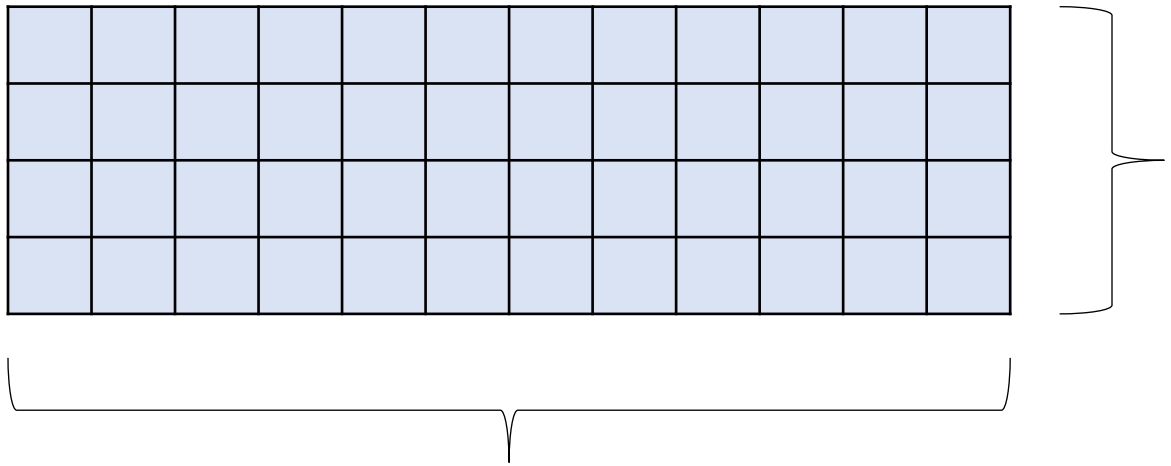


72 samples
=
72 x 22

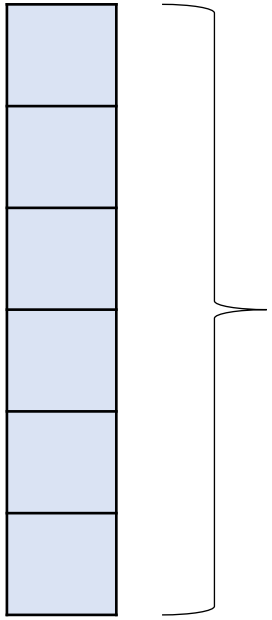


72 x (length of sample / 0.1ms)
=
7218

How about those 72 samples?



72 samples
=
72 x 22

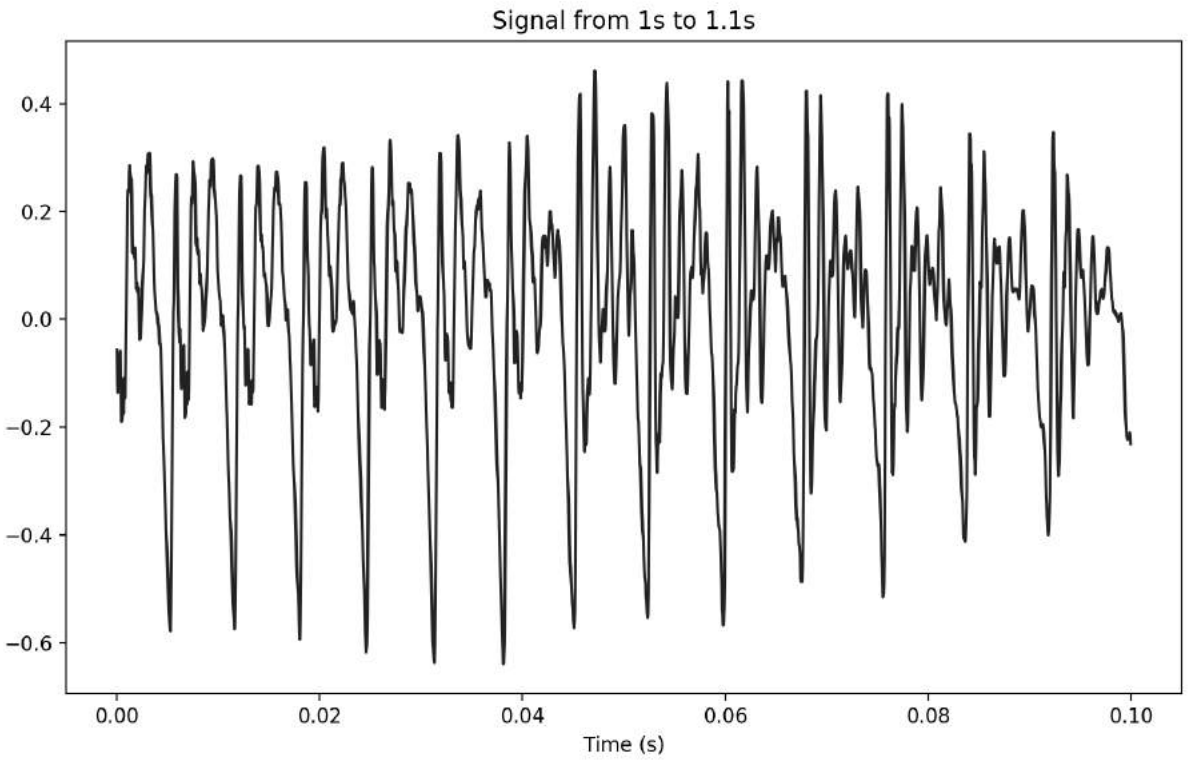


72 x (length of sample / 0.1ms)
=
7218

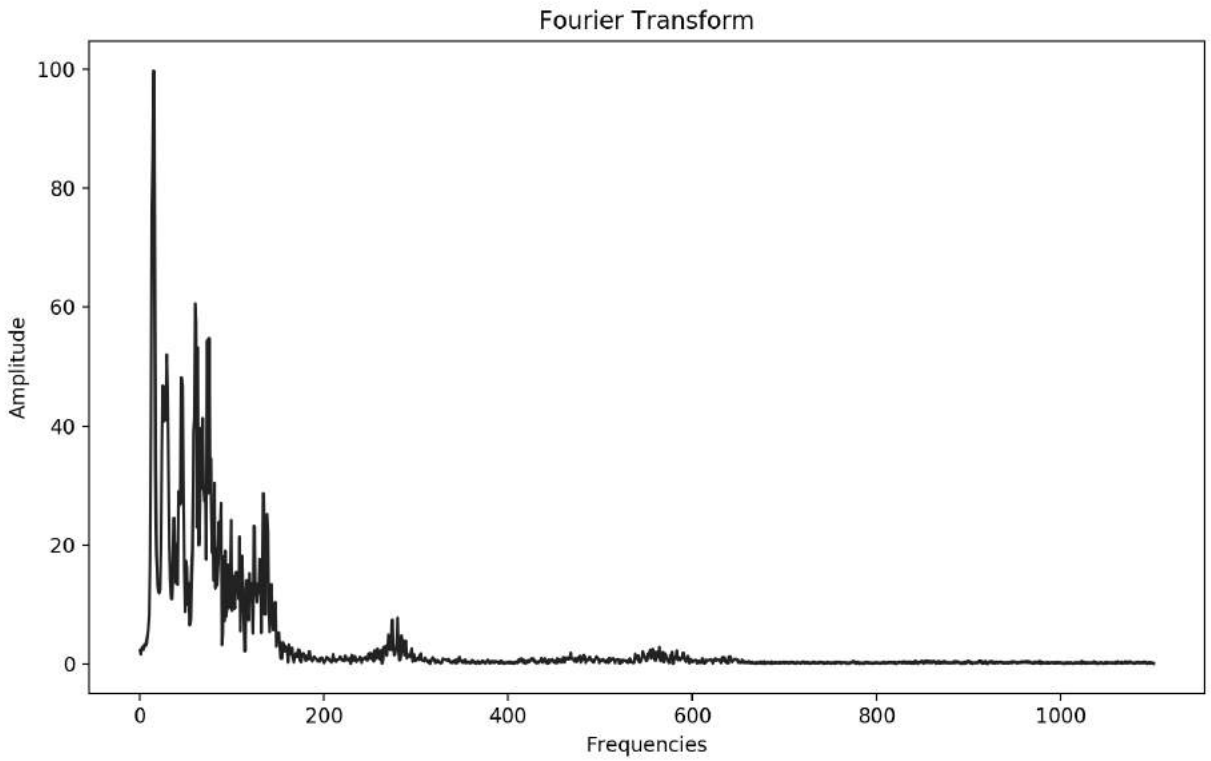
How about the signals?

Slice each file by 100ms intervals, and convert to frequency domain

Time Domain



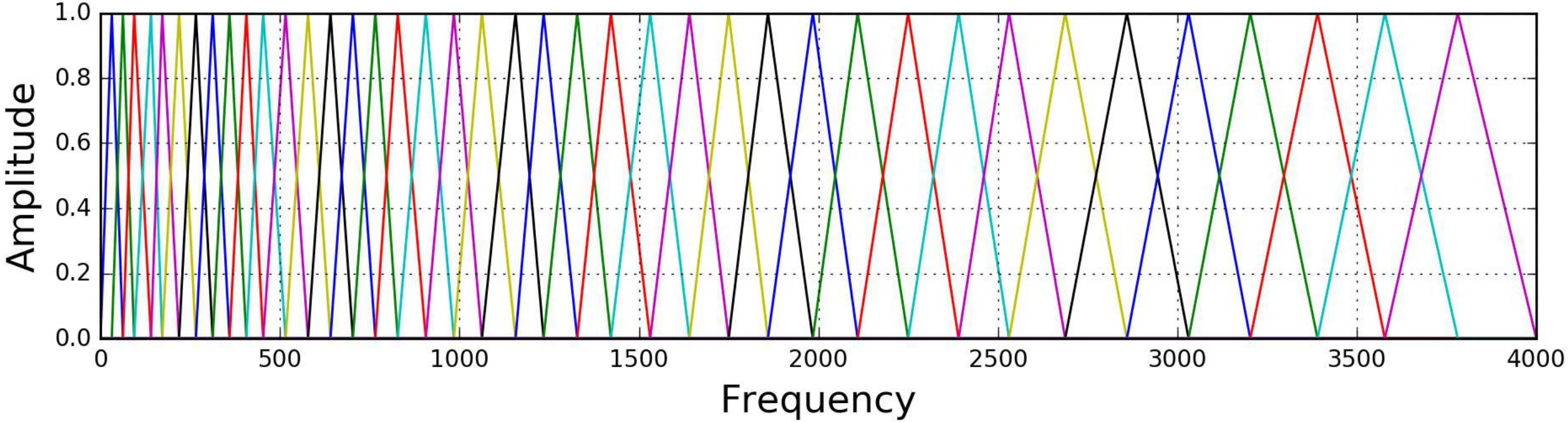
Frequency Domain



Mel Frequency Cepstral Coefficients (40 bands)

Log-scale conversion
of frequencies to
match human hearing

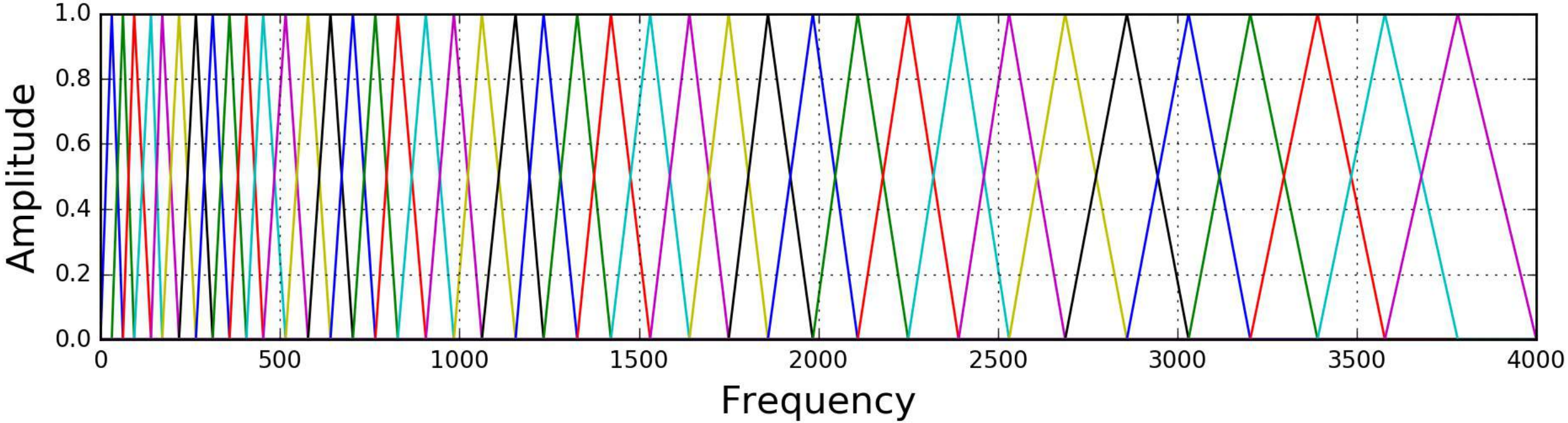
CEPSTRAL?



Mel Frequency Cepstral Coefficients (40 bands)

Log-scale conversion
of frequencies to
match human hearing

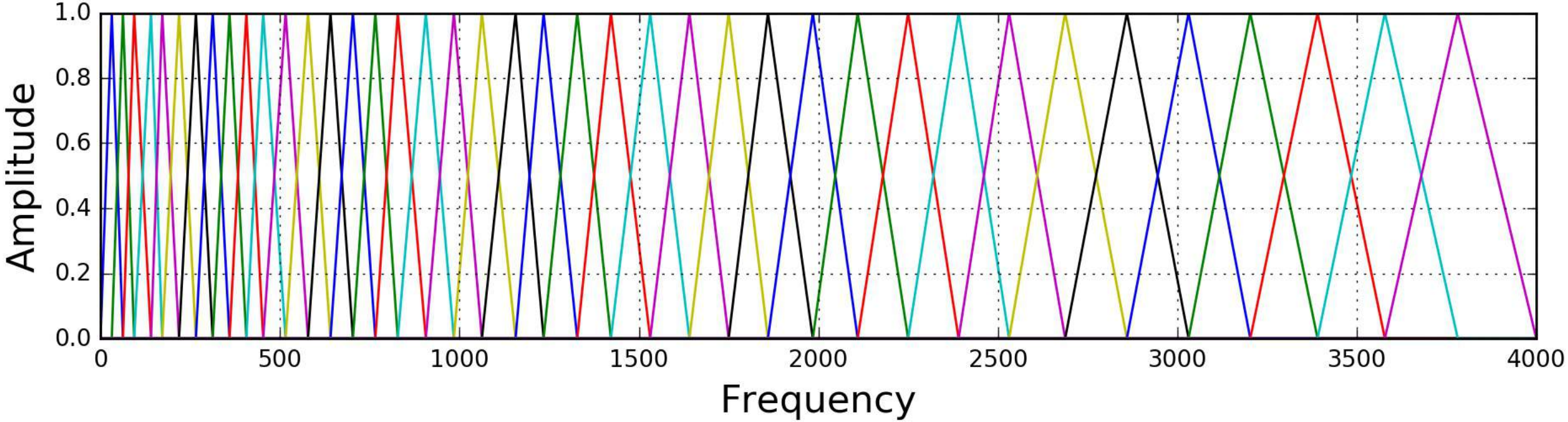
SPECTRAL!



Mel Frequency Cepstral Coefficients (40 bands)

Log-scale conversion of frequencies to match human hearing

Cepstral -> convert from frequency back to time domain.



42.44%

Significant PCC

→

58.4%

Gradient Boosting Classifier

Confusion Matrix

	English	Chinese	Korean
English	325	75	201
Chinese	108	195	162
Korean	113	82	521

Precision & Recall

	PR	RE
English	0.60	0.54
Chinese	0.55	0.42
Korean	0.59	0.73

Confusion Matrix

	English	Chinese	Korean
English	325	75	201
Chinese	108	195	162
Korean	113	82	521

Precision & Recall

	PR	RE
English	0.60	0.54
Chinese	0.55	0.42
Korean	0.59	0.73

Chinese has the lowest accuracy, possibly because it is similar to Korean.

Confusion Matrix

	English	Chinese	Korean
English	325	75	201
Chinese	108	195	162
Korean	113	82	521

Precision & Recall

	PR	RE
English	0.60	0.54
Chinese	0.55	0.42
Korean	0.59	0.73

Korean English accent is the most predictable

What else can we do?

Neural Networks
(DenseNet or CNN)

More data points

Different Preprocessing

Male-Female Split