GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of Transcribed Audio

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, et al.
gigaspeech@speechcolab.org

Outline

Why did we create GigaSpeech

How did we create GigaSpeech

Benchmark

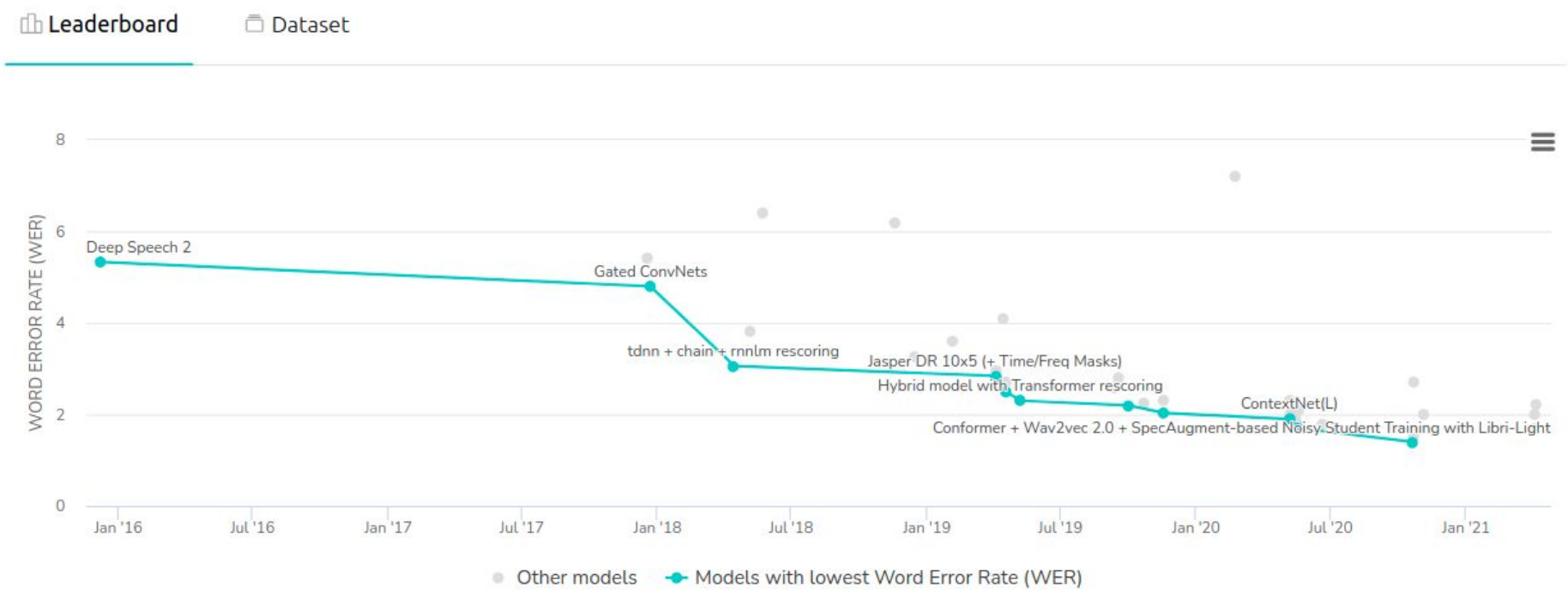
Future work

Examples

Part 1

Why did we create GigaSpeech

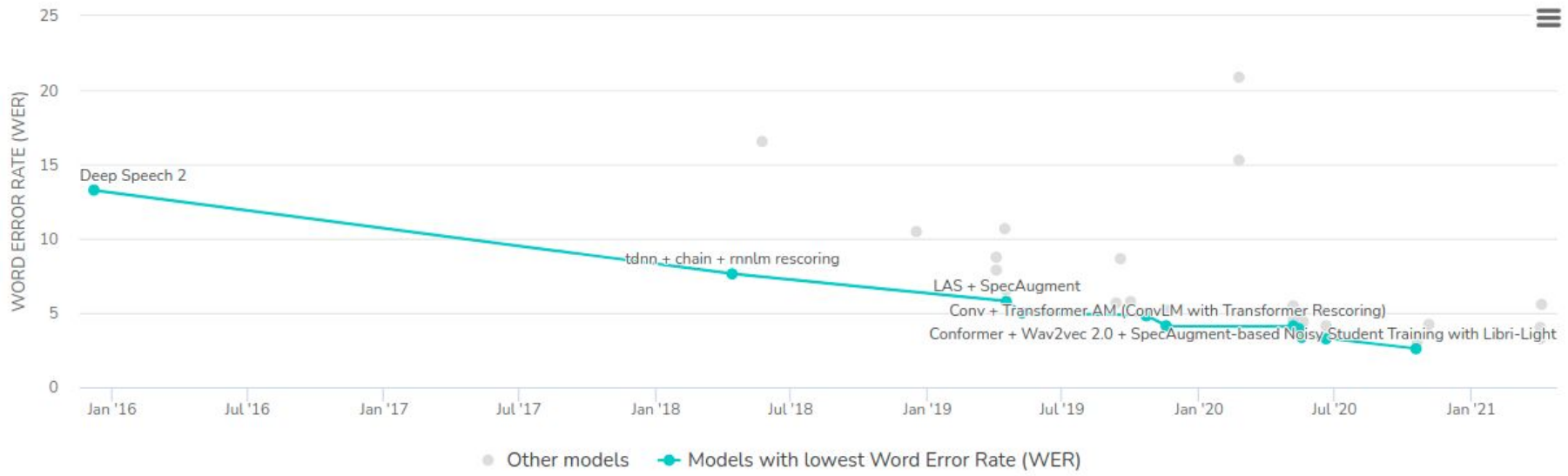
Speech Recognition on LibriSpeech test-clean



Source https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean

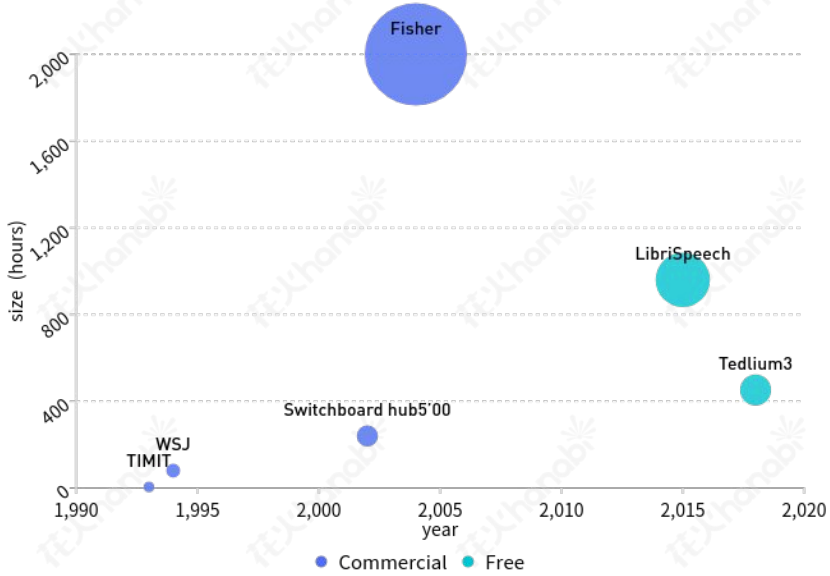
Speech Recognition on LibriSpeech test-other

Leaderboard Dataset

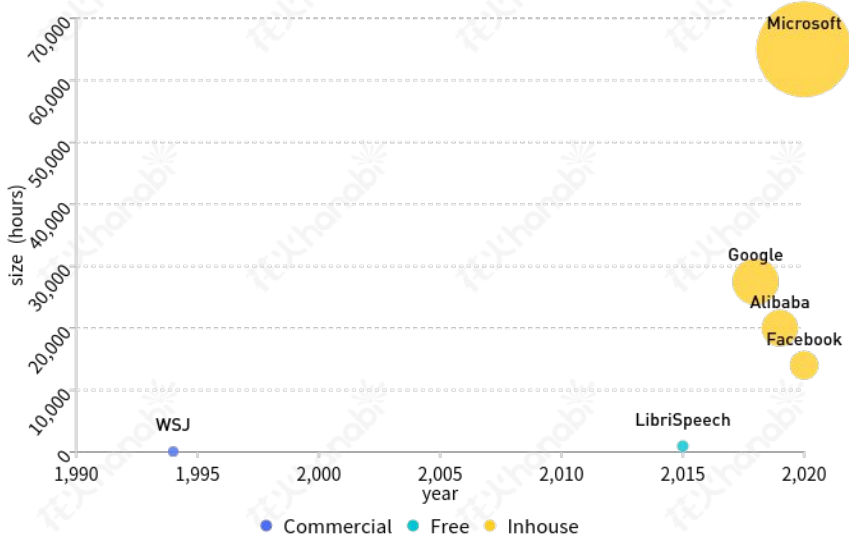


Source https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-other

Speech research datasets(academia)



Speech research datasets(industry)



GigaSpeech

Highlights

- Language: English
- 33,000+ hours for unsupervised / semi-supervised learning
- 10,000 hours with high-quality human transcriptions for supervised learning
- Multi-source, covers audiobook, podcast and YouTube
- Multi-style, covers both read and spontaneous speech
- Multi-topic, covers a variety of topics, such as arts, science, sports, etc.
- Free for research purposes

Part 2

How did we create GigaSpeech

Step 1:
Audio Collection

Audio book

- LibriVox project
- Reading speech

Podcast

- Podcast programs
- Spontaneous, Conversational speech
- Mostly indoor
- Near fields, high-quality recording
- Background musics/noises

YouTube

- YouTube website
- Various topics/speaking styles/acoustic conditions

Step 2:
Text Normalization

Standard Text normalization

- Case normalization
- Special symbol removal
- Number to word rewriting
- Date/time rewriting
- Etc.

Step 3:
Forced Alignment

Aligner

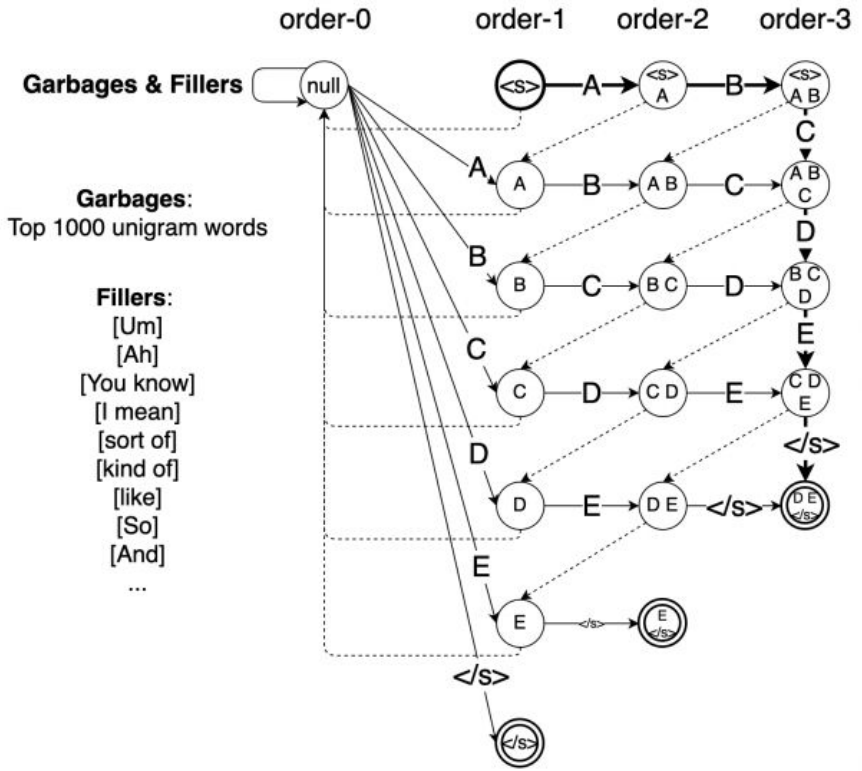
- Implemented with Kaldi
- Initial aligner trained on Librispeech
- Final aligner trained on in-domain data
- Punctuations were treated as words

Step 4:
Audio Segmentation

Segmentation rules

- Split allowed at silence that is longer than 1 second
- Split allowed at punctuation (“,”, “.”, “!” or “?”) that is longer than 0.2 seconds
- Segments with alignment WER ≥ 75% are removed
- Segments with length ≥ 20 seconds are removed
- Silence at segment boundaries are truncated to 0.15 seconds

Step 5: Segment Validation



Forced alignment graph

- State for top 1,000 unigram words
- State for filler words
- Allow insertion/deletion/substitution to the reference

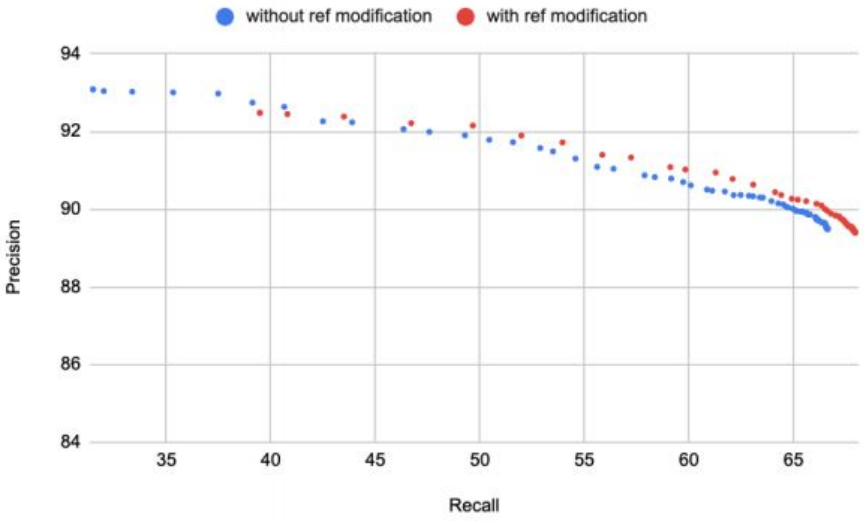
Validation decoding pass

- Detect transcription errors
- 4% maximum WER for the XL subset, and 0% for other subsets

Reference rewriting

- Allow filler words rewriting

Step 6: Evaluation



Data quality

- Frame level segmentation precision and recall on the evaluation set (manually annotated)
- Working point for the XL subset: 10,000 hours of validated data, and 4% maximum WER
- Working point for other subsets: 0% WER

Part 3

Benchmark

Evaluation Sets

Sets	Podcast	YouTube	Total
DEV	6.3h	6.2h	12.5h
TEST	16.1h	24.2h	40.3h

Creation

- Part of the evaluation set was randomly selected from the crawled data
- Part of the evaluation set was manually selected to add diversity
- Professional annotation

Diversity

- Accents, ages, Genders etc
- Acoustic conditions
- Domains, contents, topics
- Noises of all types & levels
- Speaking styles

Leaderboard

Contributor	Toolkit	Train Recipe	Train Data	Inference	Dev/Test WER
Baseline	Athena	Transformer-AED + RNNLM	GigaSpeech v1.0.0 XL	model example	13.60 / 12.70
Baseline	Espnet	Conformer/Transformer-AED	GigaSpeech v1.0.0 XL	model example	10.90 / 10.80
Baseline	Kaldi	Chain + RNNLM	GigaSpeech v1.0.0 XL	model example	14.78 / 14.84
Baseline	Pika	RNN-T	GigaSpeech v1.0.0 XL	model example	12.30 / 12.30
Mobvoi	Wenet	Joint CTC/AED(U2++)	GigaSpeech v1.0.0 XL	model example	10.70 / 10.60

Make contributions:

https://github.com/SpeechColab/GigaSpeech

Part 4

Future work

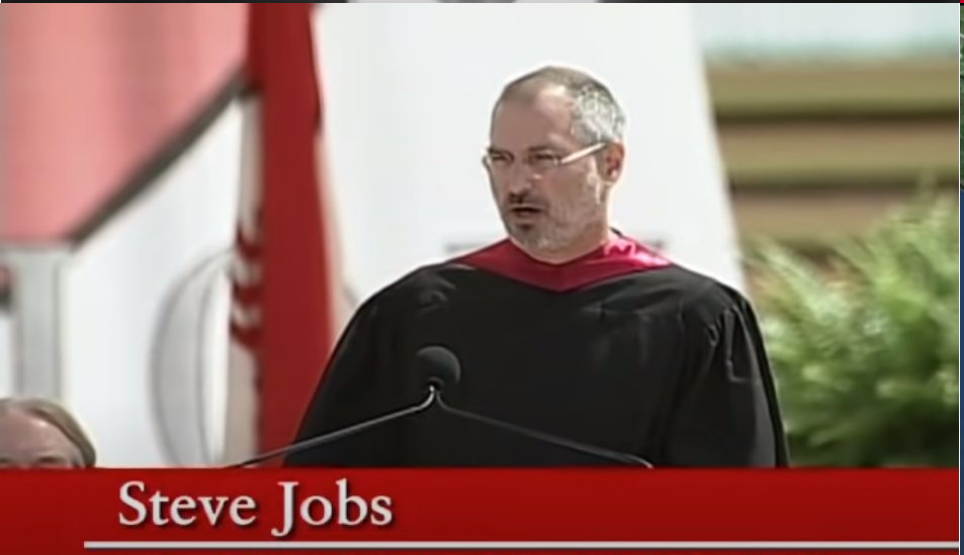
Future Work

- More languages
- More benchmarking
- Pre-trained models
- Fine-tuning
- PySpeechColab
- Decoders

Part 5

Examples


PRES. BIDEN'S
'WARTIME'
COVID PLAN
BREAKING NEWS


Steve Jobs


63
Valencia
60
Lancaster
51
Wrightwood
56
Apple Valley
64
Van Nuys
62
Pasadena
63
San Bernardino
62
Los Angeles
62
Santa Monica
64
Anaheim
65
Riverside
61
Long Beach
60
Laguna Beach
62
Temecula


HOLLYWOOD


ellentube


2006

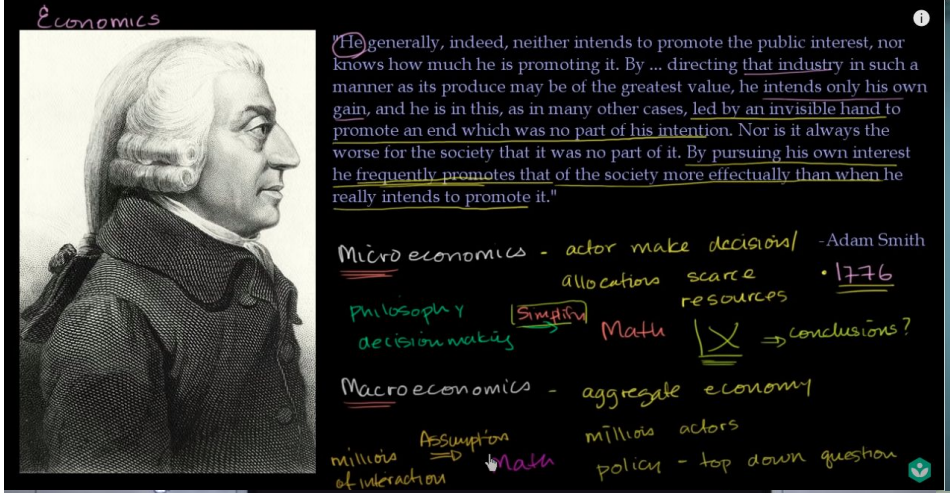

FIFA TV
HYUNDAI
HYUNDAI
HYUNDAI


TIME WARNER CABLE SPORTSNET LIVE
NBA TV
94
84


MGM GRAND
SOUTHPAW
JULY 24
UFC
MCGREGOR
UFC 3:48
MENDES

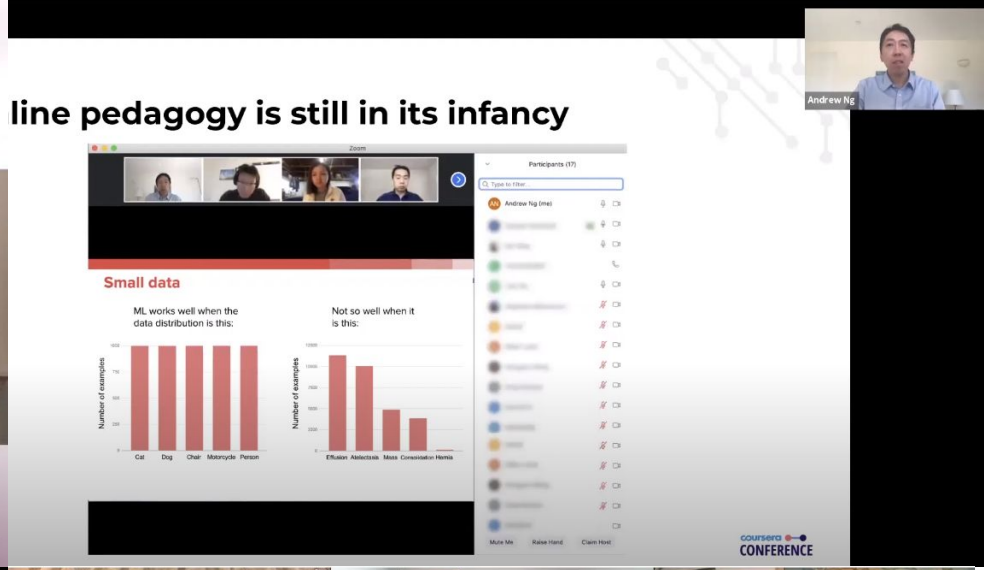

3RD & 4
3RD DOWN CONVERSIONS
1/4


LIVE
9:43 pm ET


Economics
"He generally, indeed, neither intends to promote the public interest, nor knows how much he is promoting it. By ... directing that industry in such a manner as its produce may be of the greatest value, he intends only his own gain, and he is in this, as in many other cases, led by an invisible hand to promote an end which was no part of his intention. Nor is it always the worse for the society that it was no part of it. By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it."
-Adam Smith
Microeconomics - actor make decisions/ allocations scarce resources
• 1776
Philosophy decision making
Simplify
Math
⇒ conclusions?
Macroeconomics - aggregate economy
millions of interaction
Assumption
Math
millions actors
policy - top down question


Financial Markets
Prof. Robert Shiller, Yale University


ROTHSCHILD FAMILY
OF FRANCE


line pedagogy is still in its infancy
Andrew Ng
Small data
ML works well when the data distribution is this:
Not so well when it is this:
Number of examples
Cat
Dog
Chair
Motorcycle
Person
Effusion
Atelectasis
Mass
Consolidation
Hernia
Participants (17)
Andrew Ng (me)
Mute Me
Raise Hand
Claim Host
coursera
CONFERENCE


Dr Worknoh Gebeyehu

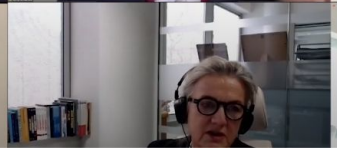

Her Excellency Amira Elfadil Mohammed Elfadil

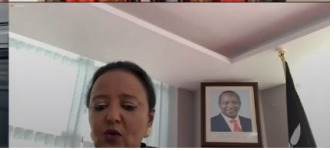

Hindou Oumarou Ibrahim


Goran Buldioski


Kamal Amakrane


Kumi Naidoo



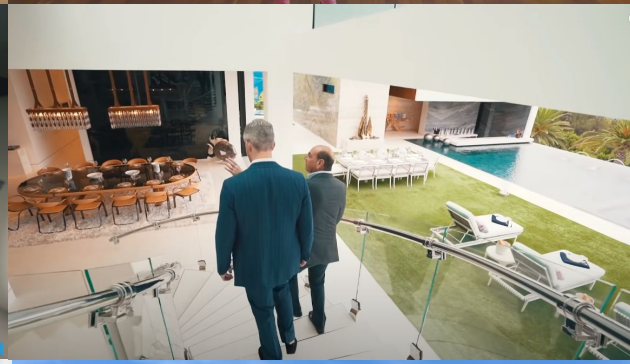


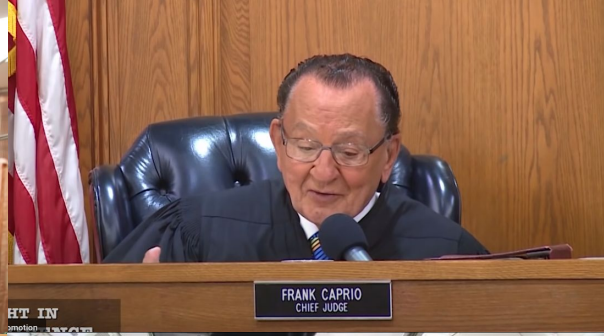

15:44
App Store
Connecting to TicWatch E2 0061...
When prompted, confirm Bluetooth pairing request.
On your phone, download and open:
Wear OS by Google


NRG Esports
1
Round 3/30
1
Natus Vincere
Best of 3 Series
INFERNO
MIRAGE
TRAIN
Ethan
$3350
CeRq
$300
tarik
$350
Brehze
$3350
gaps
$700
s1mple
Oleksandr Kostyliev
16
72
19/85
$600






carwow




FRANK CAPRIO
CHIEF JUDGE


"I've never had much luck finding trea
did once find the North Pole. Do you
like that?"
The friends thought this an excelle


JORDAN WATSON

Thank You!

Try GigaSpeech:
https://github.com/SpeechColab/GigaSpeech