ConEC:

Earnings Call Dataset with Real-world Contexts for Benchmarking Contextual Speech Recognition

LREC-COLING 2024

Ruizhe Huang[1], Mahsa Yarmohammadi[1], Jan Trmal[1], Jing Liu[2], Desh Raj[1], Leibny Paola Garcia[1], Alexei V. Ivanov[4*], Patrick Ehlen[5*], Mingzhi Yu[2], Ariya Rastrow[2], Daniel Povey[3], Sanjeev Khudanpur[1]

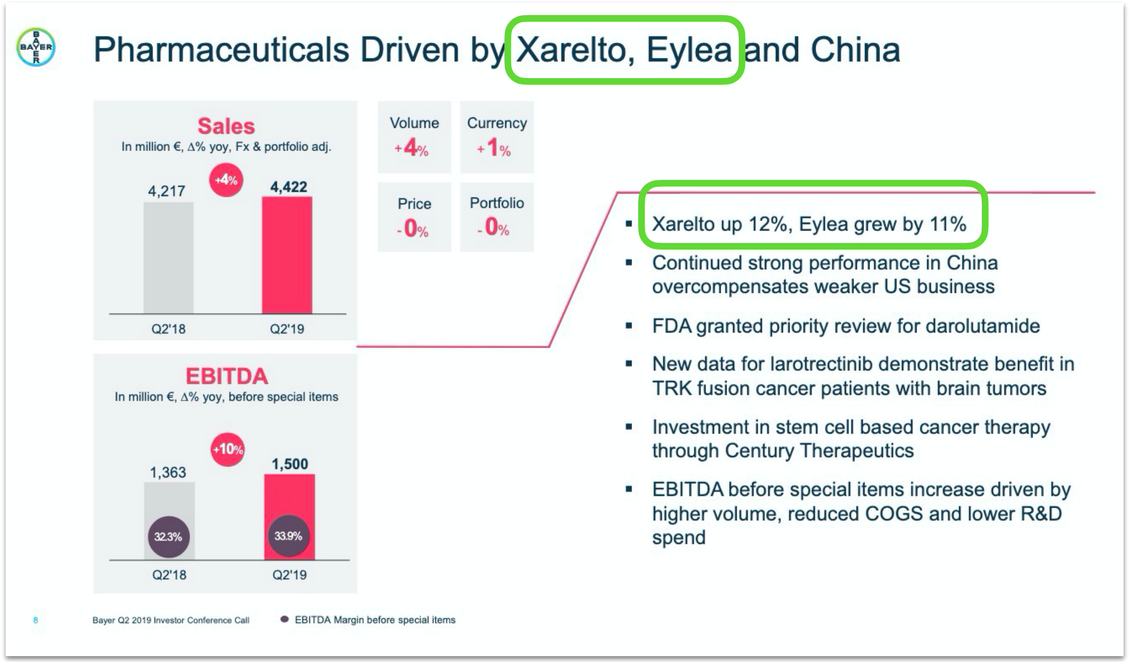
[1]Johns Hopkins University, [2]Amazon Alexa, [3]Xiaomi Corp., [3]AMD, [3]VoiceBrain

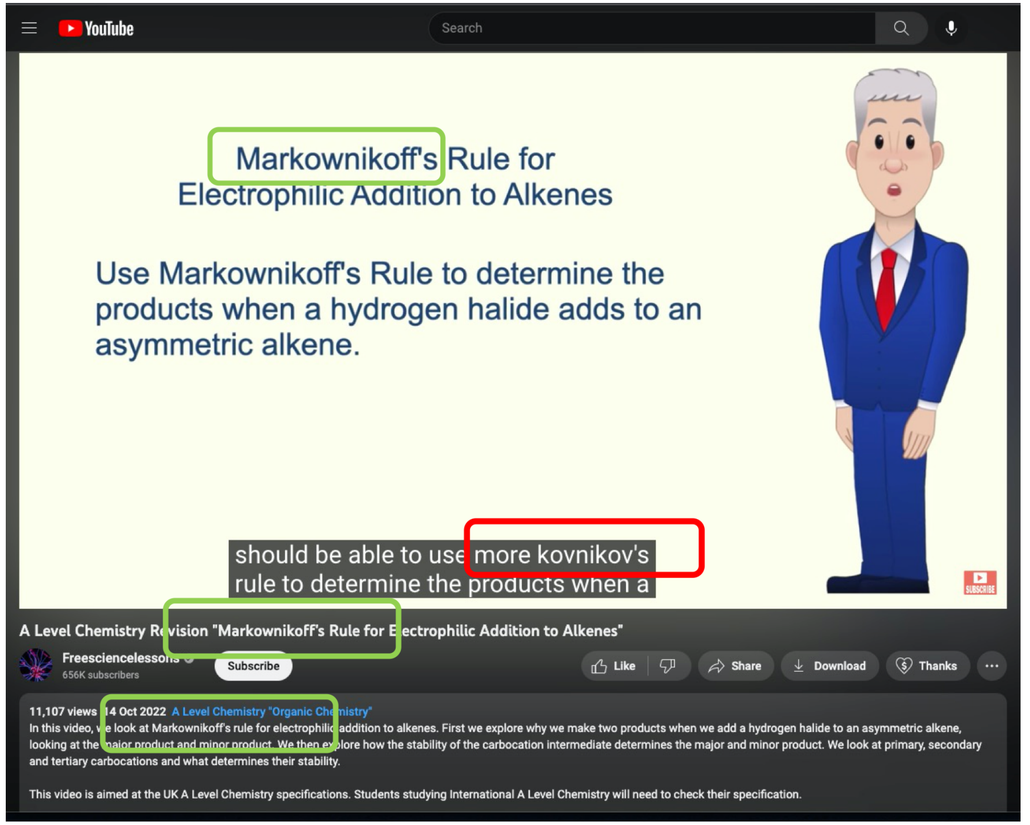
{ruizhe, mahsa, khudanpur}@jhu.edu

*: This work was done when the authors were at Uniphore.

Motivation

- Recognizing **contact names**, **proper nouns**, **rare words** and **other user-specific vocabulary** in ASR is still challenging





Ref: our 2 blockbuster drugs **eylea** and **xarelto** showed strong performance
Hyp: our 2 blockbuster drugs **alea** and **zarelto** showed strong performance

Sources:
1. Bayer AG Q1 2021 Earnings Call
2. https://www.youtube.com/watch?v=8zBLPLiXK-w&t=21s

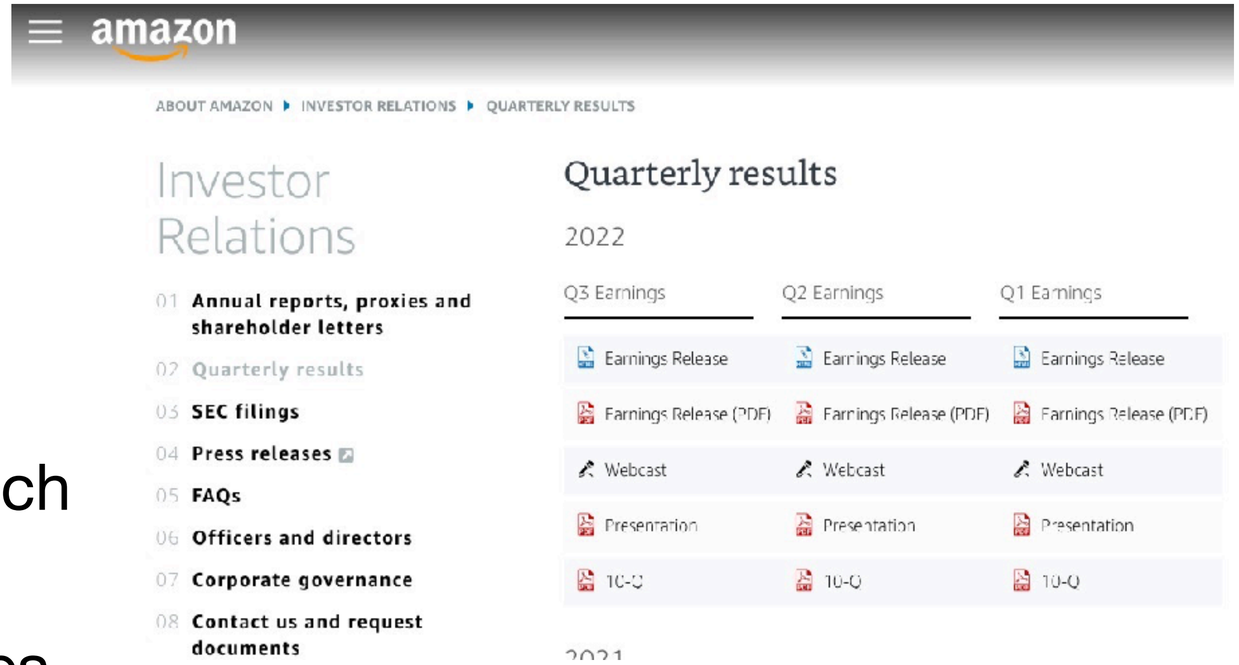
Motivation

- Can speech recognition (ASR) benefit from the available contexts?
- How to benchmark existing solutions?

Yes, and we will introduce a dataset with real-world application and several baselines for the benchmark

ConEC

- **Earnings calls** (ECs) are publicly available teleconference calls
- Executives of public companies discuss their financial results and answer questions from the participants
  - Introduction
  - Welcome and overview
  - Detailed overview
  - Q&A
- Narrated and spontaneous speech
- Contains many named entities, e.g., names of people, companies and products



ConEC

- A public corpus with **real-world context** based on public domain ECs
  - **Audio**: Earnings21/22* or Seeking Alpha
  - **Transcripts**: Earnings21/22 with our correction
  - **Slides and earnings news release**: downloaded from the companies' websites
  - **Names and affiliations** of meeting participants: Seeking Alpha
  - **Segmentation**: we provide segmentation for the data
  - **ASR baselines**: our implementation in icefall toolkit
  - **Benchmark leaderboard**: we will maintain a leaderboard on GitHub

Available on github: https://github.com/huangruizhe/ConEC

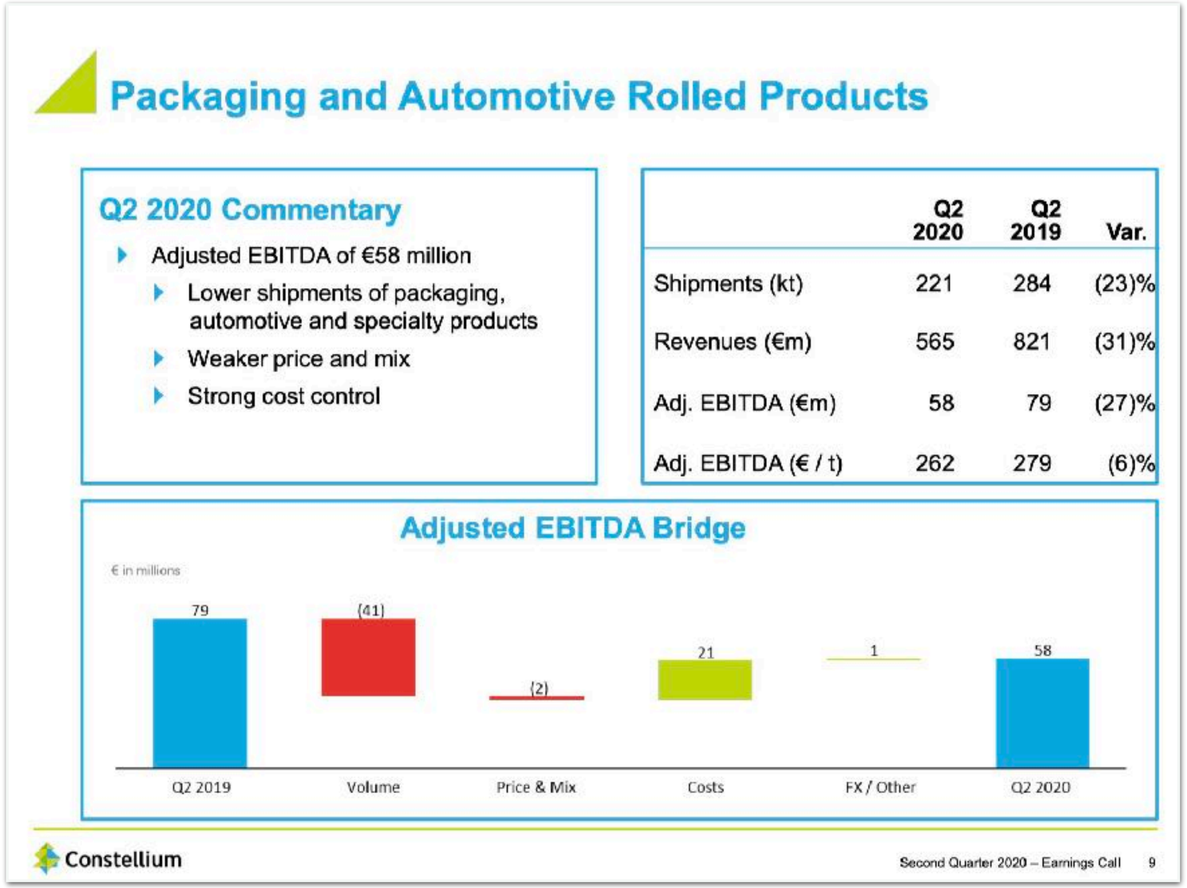
* Miguel Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, et al. Earnings-21: A practical benchmark for ASR in the wild. Interspeech 2021.
Miguel Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra. Earnings-22: A practical benchmark for accents in the wild. Interspeech 2022.

ConEC

- ConEC
  - **Earnings-21**: 44 EC recordings, 54 minutes each on average, durations range from less than 17 minutes to 1.5 hours
  - **Earnings-22:** 125 EC recordings, 57 minutes each on average. They are accented English calls from 7 regions over the world.
- The transcripts (364K and 1M tokens) are provided by human annotators and are verbatim, but it requires some correction. More on this later.
- The ECs are rich in named entities, e.g., 1.7% and 1% tokens are tagged as organization and person, which are the two most popular entity types
- Each EC has some supplemental materials (slides, earnings release, or both) depending on the availability on the website; as well as other meta information.
  - Various lengths, ranging from a few slide pages to a report of 30 pages

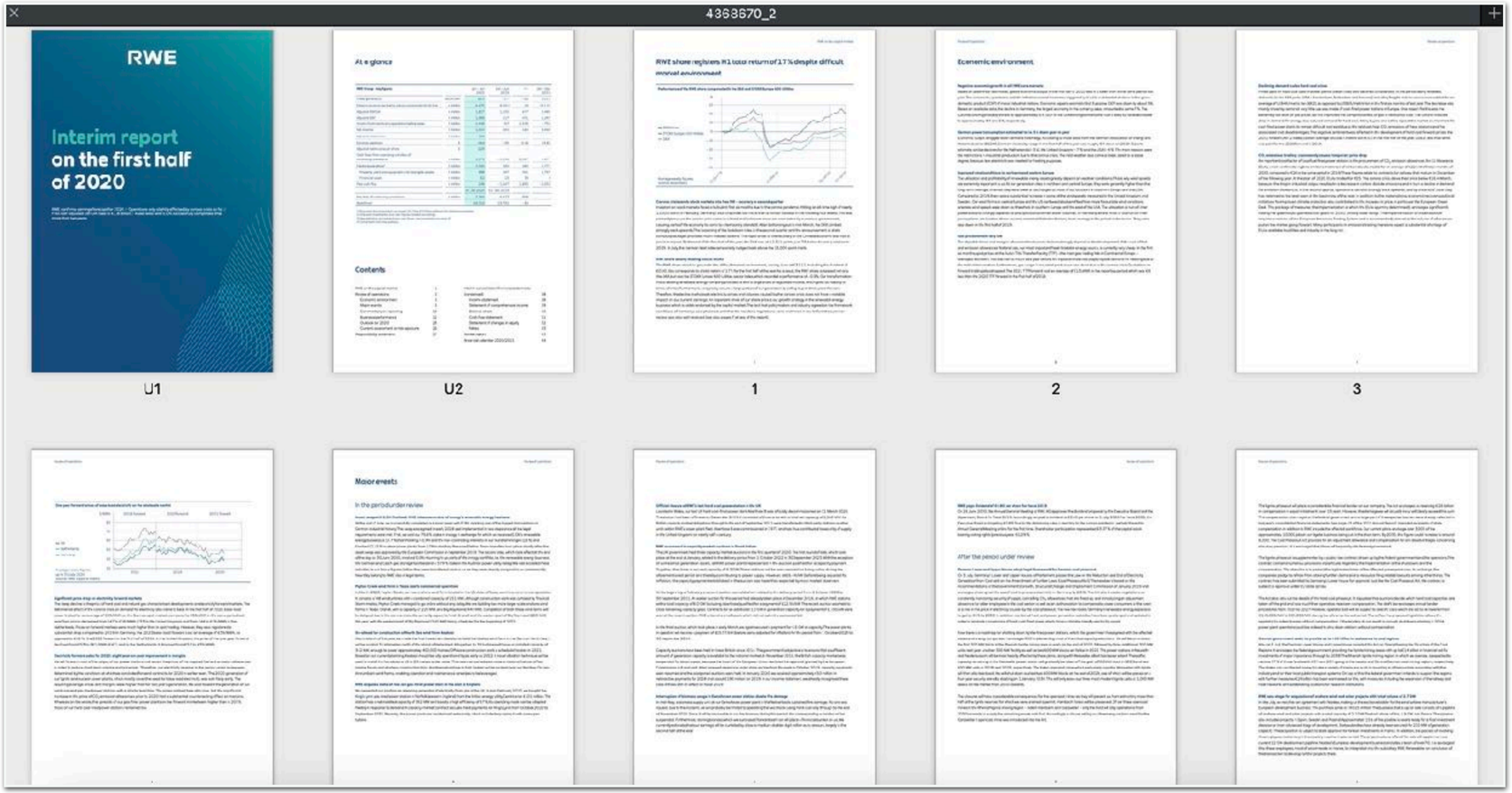
ConEC

- Example of supplemental materials



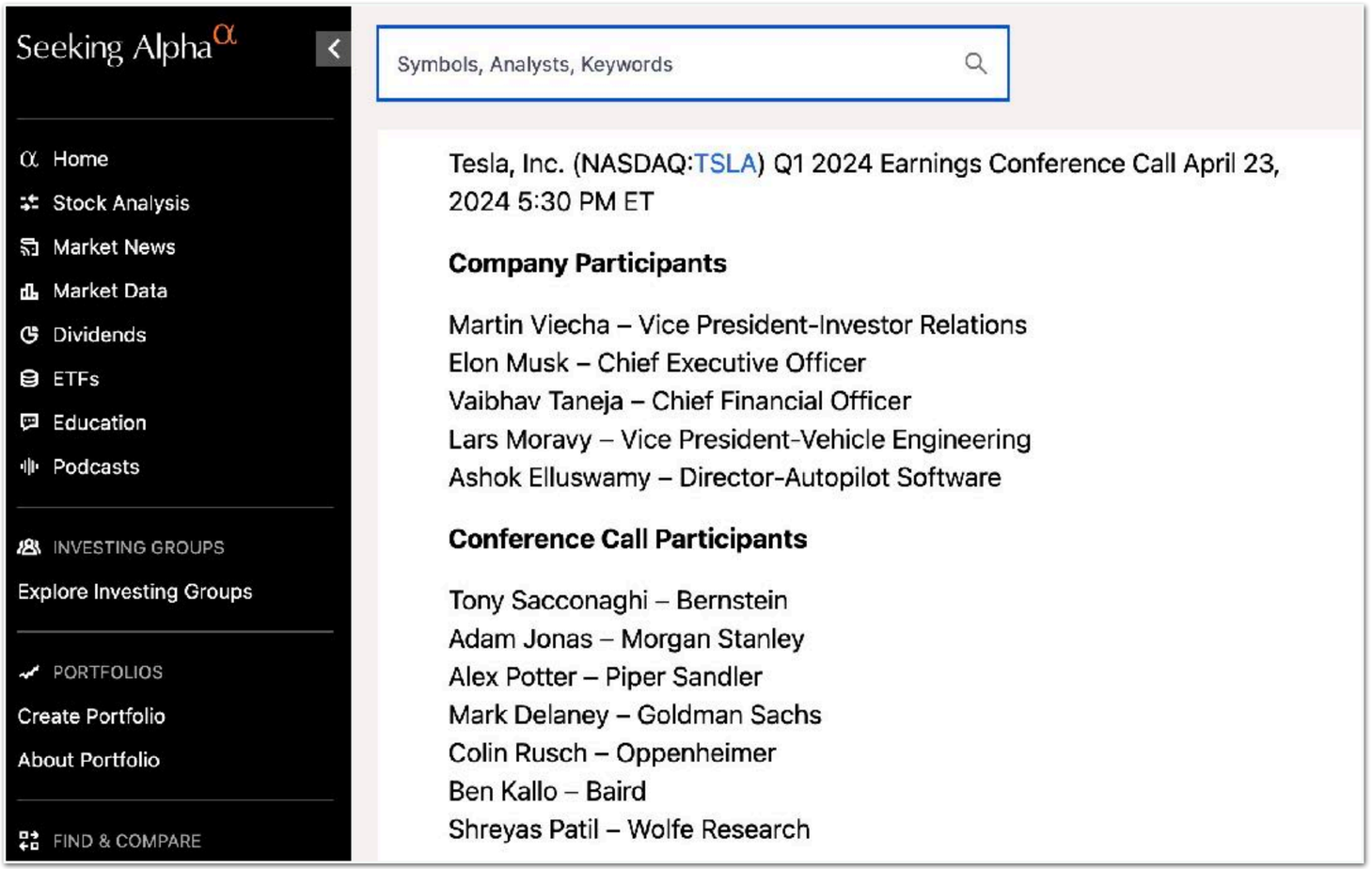
ConEC

- Example of supplemental materials



ConEC

- Example of supplemental materials



ConEC

- How much useful information can the supplemental materials provide?
- Counts of different named entity types*, and their coverage percentage by the collected context in ConEC (Earnings-21):

Entity Type	Cvrg / Count	Entity Type	Cvrg / Count
PERSON	82% / 3340	PRODUCT	39% / 671
ORG	66% / 6362	EVENT	39% / 575
GPE	61% / 1605	NORP	39% / 201
LOC	48% / 532	FAC	29% / 181

- The supplemental materials provides a good coverage for some entities, but it does not provides 100% coverage.

*: Details of entity types can be found on https://github.com/revdotcom/speech-datasets/tree/main/earnings21

ConEC

- Transcripts issue:
  - There are various sources of EC transcripts.
  - Among them, the transcripts in **Earnings21/22** are verbatim, publicly available and free. However, there are occasional misspelling errors for named entities, as well as some <unk> and <inaudible> tokens, indicating the transcription task is hard even for humans.
  - On the other hand, **S&P Capital IQ Transcripts**, which is not free, provides professionally transcribed non-verbatim text. Consistent with the real world.
  - We corrected Earnings21/22 transcripts in a semi-automatic way, by aligning it to S&P transcriptions. Heuristics and manual verification are performed.
    - Without correction, it's challenging to relate them to real-world contexts

Existing Datasets

- Existing earnings calls datasets
  - SPGISpeech (Interspeech 2021): 5000 hours
    - Highly pruned. E.g., segments containing names that appear fewer than ten times are discarded.
    - No meta data (e.g., company, year, quarter)
  - Earnings-21 and Earnings 22 (Interspeech 2021/2022)
    - The EC supplemental materials are missing
    - Occasional mistakes in transcripts, inconsistent with the real-world
  - S&P Capital IQ Transcripts (Licensed)
    - Professionally transcribed but not freely available
    - No audio.

Existing Datasets

- Existing public ASR+contexts datasets
  - AMI corpus (2005) contains 22 recorded meetings accompanied by slides in dev/test sets
  - French multimodal educational dataset of oral courses (LREC 2020) provided 10 hours of lectures where slides are available
  - LPM Dataset (ICCV 2023) has 180+ hours of video and 9000+ slides.
  - SlideSpeech (2023) provides 1,705 videos with a total duration of 1,000+ hours, and 473 hours of transcribed speech.
    - ECs belong to very different domains (vs., lectures)
    - ECs have direct application in finance industry. ECs have unique properties, e.g., time sensitivity, a lot of numeric values, various contexts, etc.
    - ECs likely contain a more diverse range of entities and unseen entities than lectures

Baselines

- Conventionally, ASR is modeled as:

$$W^* = \text{argmax}_W P(W|X)$$

- With additional contexts C, ASR can be modeled as:

$$W^* = \text{argmax}_W P(W|X, C)$$

- We choose a simple yet robust baseline — **shallow fusion**:

$$W^* = \text{argmax}_W (\log P(W|X) + \lambda \log P_C(W))$$

Baselines

- Our model is implemented in the icefall* toolkit
  - Zipformer transducer (Yao et al., 2023) of 71.5M parameters
  - Trained on SPGISpeech (5000h)
  - We treat the contexts simply as a bag-of-word model
- We also take the (unbiased) Whisper* model as another baseline (SOTA)
  - Encoder-decoder Transformer
  - Trained on 680k hours of unpublic data
  - Tiny/base/large models has 39M/74M/1550M parameters

*: https://github.com/k2-fsa/icefall/
https://github.com/openai/whisper

ASR Performance

- ConEC contexts can provide reasonable WER gain (row 4) over ASR without biasing (row 1) and a very low "oracle" WER indicating the room to improve (row 5)
- Whisper (large) is a strong baseline, even without contextual biasing (row 8)

		WER (Comm / Rare)	None	PERSON	ORG	GPE	LOC	PROD.	EVENT	NORP	FAC
No bias	1	No biasing	10.41 (8.71 / 26.02)	9.40	45.9	29.5	18.8	5.85	24.2	43.1	9.55	28.7
Shallow fusion	2	Le et al. (2021)	10.08 (8.62 / 23.43)	9.18	40.7*	25.6*	17.8*	5.26	20.2*	42.3	8.04	25.4*
	3	Fox and Delworth (2022)	10.22 (8.62 / 24.80)	9.35	38.9*	25.3*	19.2*	5.65	23.5*	41.9	10.1	29.8*
	4	***ConEC***	**10.29 (8.70 / 24.84)**	**9.39**	**39.8***	**26.1***	**18.4**	**5.65**	**21.9***	**43.1**	**9.55**	**28.7**
	5	***ConEC* (oracle)**	**9.69 (8.71 / 18.72))**	**9.25**	**13.0***	**17.7***	**12.9***	**5.46**	**19.2***	**35.6***	**5.53***	**16.6***
Whisper	6	Whisper (tiny)	19.16 (16.79 / 40.62)	18.4	61.1	47.8	31.1	18.9	46.1	8.00*	25.1	54.1
	7	Whisper (base)	14.67 (12.72 / 32.37)	13.9	51.5	40.1	25.9	14.8	35.5	7.30*	18.1	48.1
	8	Whisper (large)	7.98 (6.94/17.43)	7.50	28.9*	19.6*	17.0*	5.85	18.7*	2.78*	8.54*	21.6*

Synthesized

Real-world

WERs on Earnings-21. * means statistically significant improvements over no biasing.

Challenge and Future Work

- We introduced ConEC, a corpus and a benchmark for contextualized ASR grounded on real-world contexts (e.g., slides) and application
- The contexts are challenging yet still provide good coverage for named entities and help improving ASR
- There are several unexplored perspectives and open research opportunities of ConEC for the future:
  - As LLMs becomes popular, can LLMs consume the slides directly instead of using a bag-of-words model?
  - We haven’t investigated the numeric values in slides and financial reports, however, they are high-stake contents
  - The Q&A part of ECs contains spoken interaction between people. Can they be better (automatically) analyzed given the contexts?

Thank you!

Check out our dataset and benchmark at:

https://github.com/huangruizhe/ConEC

If you are interested or have any questions about our work, please contact us at:
{ruizhe, mahsa, khudanpur}@jhu.edu