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

Ruizhe Huang[1], **Mahsa Yarmohammadi**[1], **Jan Trmal**[1], **Jing Liu**[2], **Desh Raj**[1], **Leibny Paola Garcia**[1], **Alexei 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., [4] AMD, [5]VoiceBrain,
{ruizhe,mahsa,khudanpur}@jhu.edu

Abstract

Knowing the particular context associated with a conversation can help improving the performance of an automatic speech recognition (ASR) system. For example, if we are provided with a list of in-context words or phrases — such as the speaker's contacts or recent song playlists — during inference, we can bias the recognition process towards this list. There are many works addressing contextual ASR; however, there is few publicly available real benchmark for evaluation, making it difficult to compare different solutions. To this end, we provide a corpus ("*ConEC*") and baselines to evaluate contextual ASR approaches, grounded on real-world applications. The *ConEC* corpus is based on public-domain earnings calls (ECs) and associated supplementary materials, such as presentation slides, earnings news release as well as a list of meeting participants' names and affiliations. We demonstrate that such real contexts are noisier than artificially synthesized contexts that contain the ground truth, yet they still make great room for future improvement of contextual ASR technology.

Keywords: speech recognition, benchmark, contextual ASR, earnings call, named entities

1. Introduction

While significant progress has been made in ASR technology in recent years, recognizing words that are not frequently seen in the training data (i.e. rare words) or named entities such as proper nouns or user-specific vocabulary is still a challenge. For example, word error rate (WER) on LibriSpeech (Panayotov et al., 2015) *test-other* set can be as low as 5%, but word error rate on rare words remains over 20% (Section 4.2; Huang et al. (2023)). This is due to the long-tailed data imbalance problem of neural network models (Zhang et al., 2021), which are usually biased towards dominant classes in the training data and cannot generalize well beyond that.

Contextual ASR aims to improve the accuracy on rare words or named entities by incorporating contextual information in addition to input acoustic signals during the recognition process. The context can come from a variety of sources such as external knowledge (Le et al., 2021; Wang et al., 2022), prior utterances (Wei et al., 2021; Yang et al., 2023), pronunciations (Pandey et al., 2023), audio or video metadata (Liu et al., 2020; Ray et al., 2021), visual contexts (Pramanick and Sarkar, 2022; Li et al., 2023), and so on. Contextual information can be incorporated into the model either in a post-training manner by shallow fusion (Zhao et al., 2019; Wang et al., 2023b) or on-the-fly rescoring (Yang et al., 2021), or at a deeper level, as an additional input during training (Pundak et al., 2018; Jain et al., 2020; Chang et al., 2021; Sathyendra et al., 2022; Yang et al., 2023; Lu et al., 2023).

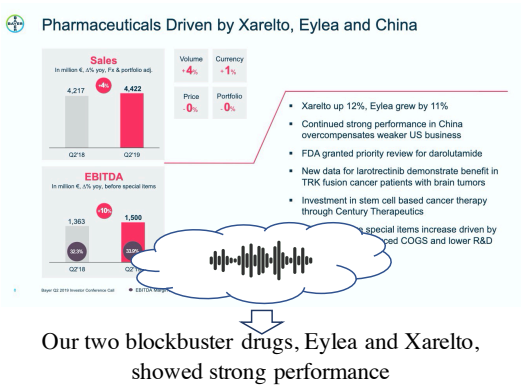

Figure 1: Real-world contexts (e.g., slides) are provided to ASR to better recognize spoken contents.

Besides variety of context types and models, various datasets have been used for evaluating the effect of context in ASR. These datasets and contexts are either synthesized that often include the ground truth (e.g., (Le et al., 2021)) or in-house data which is not publicly available (e.g., voice assistant traffic (Chang et al., 2021)). There is few publicly available real benchmark for evaluating contextual ASR, making it difficult to compare different systems. Such a benchmark should reflect real-world complexities, that are more challenging than laboratory settings. To bridge this gap, we propose a public and practical corpus ("*ConEC*") with real-world context based on public-domain earnings calls (ECs).

*This work was done when the authors were at Uniphore

LREC-COLING 2024, pages 3700–3706

Better EC transcriptions are particularly appealing because of their merit in automated financial information retrieval and analysis, e.g., by BloombergGPT (Wu et al., 2023). ECs contain a large amount of real-world named entities, such as the names of companies, products, events and people, which are hard for ASR to recognize correctly. ECs usually come with supplementary materials such as presentation slides and earnings news releases (Figure 1), providing useful contexts. This work offers:

- A contextual ASR corpus *ConEC*[1], which associates existing EC datasets with real context collected from supplementary materials.
- A public shallow fusion ASR baseline that takes into account the available contexts for improving entities and rare words.

2. Related Datasets

There are several existing corpora of EC data. Qin and Yang (2019) selected only the sentences made by the most spoken executive (usually the CEO) from 576 ECs for predicting stock's risks. SPGISpeech (O'Neill et al., 2021) contains 5,000 hours of EC data purposed for training ASR to generate fully formatted text. SPGISpeech can be used to train a reasonable ASR model in the EC domain. However, it is significantly pruned, e.g., segments containing currency information or names that appear fewer than ten times are discarded. Besides, there is no meta data (e.g., company, year, quarter) provided for each utterance, which results in a lack of context of the entire EC. S&P Capital IQ Transcripts[2] provides professionally transcribed non-verbatim text for ECs since 2002, but it does not include audio and is not freely available.

Earnings-21 (Rio et al., 2021) and *Earnings-22* (Rio et al., 2022) preserve long-form ECs. They are test sets of 39 and 119 hours, respectively. *Earnings-21* includes named entity labels that are derived from an automatic named entity recognizer and subsequently validated through manual review. In this study, we focus on non-numeric entities, as shown in Table 1. To facilitate future research, we augment *Earnings-21/22* with real-world contexts and examine their impacts on ASR.

There are several multimodal corpora that contain text, image, or video contexts accompanied by speech. AMI corpus (McCowan et al., 2005) contains 22 meetings accompanied by slides in dev/test sets, which is used by Sun et al. (2022) for contextual ASR. Mdhaffar et al. (2020) provided 10 hours of lectures, which are manually transcribed and segmented. They also come with video recordings and presentation slides. More recently, Lee et al. (2023) provided a dataset containing aligned slides and spoken language, for 180+ hours of video and 9000+ slides. Similarly, Slidespeech (Wang et al., 2023a) is another audio-visual corpus enriched with slides.

We propose to create a contextual ASR corpus based on ECs, in addition to the above datasets that belong to very different domains (e.g., lectures). Having a contextual ASR benchmark on ECs is appealing because of their direct application in finance industry. The unique properties of ECs (e.g., time sensitivity, a lot of numeric values, various contexts, etc.) make them highly suitable for contextual ASR. Additionally, ECs likely contain a more diverse range of entities and unseen entities than lectures or AMI meetings.

3. Proposed Dataset and Benchmark

The proposed contextual ASR corpus *ConEC* is based on existing *Earnings-21/22* audio and transcripts, with additional real-world contexts including (1) presentation slides and earnings release downloaded from the investor relations pages on the companies' websites[3]; (2) names and affiliations of meeting participants collected from Seeking Alpha[4]. We take *Earnings-21* as the evaluation set, while *Earnings-22* can be used as training and development set in addition to any other data excluding *Earnings-21*, e.g., SPGISpeech. As many existing ASR systems do not naturally handle long-form audios, we also provide accurate segmentation for the datasets. We will discuss data preparation details and statistics of the contexts in the following subsections. Then, we provide a simple yet effective baseline to demonstrate that the proposed contexts are noisier and more challenging than artificially synthesized contexts but still helpful for ASR.

3.1. Data Collection and Preparation

Audios and transcripts are taken from *Earnings-21/22*. *Earnings-21* consists of 44 EC recordings (files) of the average length of 54 minutes. The calls range in length from less than 17 minutes to 1 hour and 34 minutes. *Earnings-22* consists of about 119 hours of accented English calls from 7 regions in 125 recordings (files). The transcripts are provided by human annotators and are verbatim, which is different from the non-verbatim transcripts in S&P. Nevertheless, we noticed there are some `<unk>` and `<inaudible>` tokens in *Earnings-21/22*, which correspond to the names of companies or people, according to the S&P transcription for the same EC. There are also occa-

[1] Available at https://github.com/huangruizhe/ConEC
[2] https://www.capitaliq.com/
[3] E.g., https://ir.aboutamazon.com/quarterly-results
[4] https://seekingalpha.com/

sional misspelling errors for named entities, indicating the transcription task is hard even for humans. We replace the <unk> and <inaudible> tokens as well as the misspellings with the actual values from the S&P transcription in a semi-automatic way, by aligning both transcriptions based on Levenshtein distance.

Then, we segment the audio according to sentence boundaries in the transcription. We feed long audios into Gentle (Ochshorn and Hawkins, 2017) and WhisperX (Bain et al., 2023) to propose candidate start and end time for each sentence. However, these proposed timestamps are not always accurate. For example, many sentences in ECs end with a percentage, e.g., "Our revenue has increased by 21%" where 21% is an out-of-vocabulary (OOV) word and is pronounced in its spoken form as "twenty-one percent". Both Gentle and WhisperX aligners often struggle with such OOV words, resulting in improper segmentation.Thus, instead of the proposed end time for a sentence, we use the start time of the subsequent sentence to indicate the end time of the current sentence. We further feed the audio segments into Whisper to validate there is no insertion and deletion errors, as measured by computing WER, at the beginning and end of the sentences. In this way, the timestamps will be iteratively adjusted to achieve best ASR results.

The supplementary materials for each earnings call are in PDF format. We use pdf2txt tool[5] to extract the text from these PDFs. The extracted content contains noise. To improve the quality we remove stop words, non-alphanumeric words, and numeric values. We also remove the PDF pages containing the legal disclaimers. In our experiments, we take the cleaned contents as a bag of uncased, unigram words. However, people can use the supplementary materials in any way they favor, e.g., using them as n-grams, sentences or visual contexts. We also include names and affiliation information of meeting participants we collected from Seeking Alpha, as they are practical, easy-to-obtain contexts in the real world.

3.2. Dataset Statistics

The context we obtained is noisier than synthesized context that contains the ground truth transcriptions in two aspects. The real context contains only a limited coverage of named entities that are mentioned in the audios. On the other hand, it can sometimes provide information (distractors for ASR) that confuse or mislead the ASR system. For example, the "Aviation and *Renewables*" business is mentioned several times in the 2020 Q1 earnings call of General Electric (GE), but it does not show up in the slides. Instead, "*renewable* energy" appears multiple times in the slides. This kind of interference is more challenging than existing synthesized contexts, e.g., Fox and Delworth (2022) extract all PERSON and ORG entities (i.e., a full coverage) from the ground-truth transcription of *Earnings-21* and add names of Fortune 500 companies and famous CEOs as distractors. We argue that such distractors may be too simple.

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

Table 1: Counts of different named entity types (check out Appendix B or here[6]), and their coverage percentage by the collected context in *ConEC*.

The collected context for each EC recording consists of about 100 to 2000 words. Table 1 shows the counts of the named entities spoken in the audio, and what percentage of them (measured in tokens instead of types) are covered in *ConEC* contexts. For example, 3340 tokens out of 364k total tokens are labeled as PERSON[6], and 82% of them are present in the contexts. Similar to SPGISpeech, we omit all numeric entities for the time being, and leave number normalization as a future work.

From Table 1, PERSON and ORG have the highest coverage after including names and affiliation information. Otherwise, the coverage is only 30% and 56% for PERSON and ORG from the slides and earnings releases.

3.3. Baselines for Contextual ASR

We provide a public implementation of a baseline for contextual ASR in the open source toolkit icefall[7]. We decide to choose shallow fusion (Zhao et al., 2019) as the baseline, because it is a zero-shot method, which requires no training and can be combined with any ASR system. The idea is to give some bonus to the hypotheses hitting words in the biasing list during beam search. More specifically, the following scoring function is used for decoding:

$$W^* = \arg\max_W \log P(W|X) + \lambda \log P_C(W) \quad (1)$$

where W, X, C, λ are the hypothesis, acoustic features, contexts and a hyper-parameter controlling the biasing strength. $P(W|X)$ is the conventional ASR modeling, and $P_C(W)$ is a biasing scoring function that prefers hypotheses containing more words from the biasing list C.

[5] https://github.com/jalan/pdftotext

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

[7] https://github.com/k2-fsa/icefall

		WER (Comm / Rare)	None	PERSON	ORG	GPE	LOC	PROD.	EVENT	NORP	FAC
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
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***
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*

Table 2: WER (%), common/rare words WER and non-entity/entity WER on *Earnings-21* with different contexts. The results using *ConEC* context are highlighted. Results with * mean the improvement over row 1 is statistically significant with p<0.05 (Appendix A).

On the other hand, we compare our baseline with Whisper (Radford et al., 2022), which is a transformer sequence-to-sequence ASR system trained on 680k hours of supervised data collected from the web. The large-scale/multi-task training give Whisper impressive robustness and generalization ability, even for some long-tailed data. We simply evaluate Whisper on *Earnings-21*, without any contextual biasing, i.e., decoding with $P(W|X)$ only.

4. Experiment Results

4.1. Experiment Setup

Our ASR model is a zipformer transducer (Yao et al., 2023) of 71.5M parameters, implemented with the icefall toolkit. It is trained on normalized SPGISpeech, i.e., ignoring punctuation and letter cases. We use the fstalign[8] tool to evaluate overall WER, WER for common/rare words, and WER for each entity types. An ideal system should reduce rare words and entities WER without hurting common words WER. Hypothesis and reference texts are normalized with Whisper normalizer[9] before being fed into fstalign.

To validate our shallow fusion baseline, we also report results on Librispeech data and synthesized biasing lists (Le et al., 2021) of rare words.

4.2. Results

Librispeech. On Librispeech (Table 3), the WER without contextual biasing on *test-other* is 5.22 overall, and 3.32/21.83 for common/rare words. Simple shallow fusion method with the synthesized context consisting of all rare words in the reference transcription combined with 500 distractors for each utterance, remarkably reduces the rare words WER by 50% without harming common words WER. Our results even outperform a fully neural network based approach (Yang et al., 2023). This validates our implementation, and also indicates that synthesized context may not be sufficient to draw insights when comparing alternative contextual ASR approaches, as neural network based approach is reported outperforming shallow fusion in several existing works (Pundak et al., 2018; Le et al., 2021; Chang et al., 2021; Sathyendra et al., 2022).

	test-clean			test-other		
	All	Com	Rare	All	Com	Rare
No biasing	2.17	1.25	9.65	5.22	3.32	21.83
Shallow fusion	1.59	1.26	4.22	4.11	3.32	11.05
Yang et al. (2023)	2.00	-	-	4.45	-	-

Table 3: Librispeech WER (%) breakdown (all, common & rare words) for contextual ASR baselines.

ConEC. Next, we examine how the proposed real-world contexts in *ConEC* help contextual ASR, compared to synthesized contexts. Table 2 row 1 shows the WER breakdown of our ASR model trained on SPGISpeech without any contextual biasing. We define rare words to be the words with frequency below top-3k in the SPGISpeech training text, which accounts for 10% of total tokens. Around 70% of named entities are rare words. Table 2 demonstrates that both unbiased and biased systems face greater difficulty in recognizing rare words compared to common words. Furthermore, in both systems, the majority of named entities are more challenging to recognize compared to none-entities.

Synthesized Context. Table 2 row 2 uses synthesized contexts obtained in the way described in Le et al. (2021). For each utterance, it extracts rare words from reference transcription and adds 500 other random rare words as distractors. The resulting rare words and entity WERs are reduced to different extents, compared to row 1. The common words and non-entity WERs are reduced too. In Row 3, synthesized contexts (Fox and Delworth, 2022) are obtained as described in Section 3.2. Only PERSON and ORG entity words and distractors are included into one single biasing list of 3066 unigram words, shared by all ECs. As the result, the WERs for PERSON and ORG are reduced. The WERs for some other entity types and com-

[8] https://github.com/revdotcom/fstalign

[9] https://github.com/kurianbenoy/whisper_normalizer

mon words get worse, due to the shared biasing list containing many distractors.

Real-world Context. Table 2 row 4 (*ConEC*) uses our per-EC biasing lists extracted from slides, earnings release and participants' names and affiliations. With such contexts, we get WER reductions for rare and entity words, however, the reductions are smaller compared to row 2. We also observe smaller WER reduction for PERSON and ORG classes compared to row 3. On the other hand, common words and non-entity words have higher WERs, which indicates such real-world contexts are noisier. We speculate that noisy contexts have higher impact in training of the neural-based contextual ASR. Indeed, the distractors here are more relevant to the context in general and, as such, may represent a confusing plausibility. To assess the full potentials of real-world contexts, we report the oracle WER in row 5, based on the ASR results in row 1. More specifically, we take the ASR output of row 1 and replace the entity words with the correct ones, if they are found in the contexts. This establishes a lower bound of entity WER for our ASR model with the given contexts. There is still a significant gap between actual (row 4) and oracle (row 5) WERs, which hopefully will be filled by future contextual ASR work.

Whisper baselines. Finally, we provide Whisper's ASR output (without contextual biasing) for reference. The Whisper tiny, base, large models has 39M, 74M, 1550M parameters, where Whisper base has similar size as our model (71.5M). Note that Whisper base does not outperform our model (row 1) on this test set, even though it is trained on 680k hours of data. Whisper large has the lowest overall WERs, but recognizing some entity types that require specific knowledge (e.g., PERSON, ORG, GPE, PRODUCT, FAC) remains challenging with WERs greater than 15%.

5. Conclusion

We provide a corpus and a benchmark for contextualized ASR grounded on real-world application – transcribing earnings calls with their supplementary materials, including presentation slides, earnings news releases and meeting participants' names and affiliations. Such contexts are noisier yet still provide reasonable coverage for named entities that are hard to recognize by existing ASR systems. Along with the corpus, we also release a public shallow fusion contextual ASR baseline implemented in an open-source ASR toolkit.

We believe ConEC corpus is valuable resource for development and evaluation of the rich transcription systems, which can gain a profound understanding (vs. a bag of word model) of the available contexts by itself and facilitate automated analysis of the challenging spoken interaction between people (e.g., the "Questions and Answers" portion of each earning call contains a recording of real, high-stake spoken interactions between participants potentially under significant amount of stress).

6. Acknowledgements

We thanks the reviewers for their valuable comments. We thank Aravind Ganapathiraju, Pavankumar Dubagunta, and Grant Strimel for helpful discussions. Our work was supported by a fellowship from JHU + Amazon Initiative for Interactive AI (AI2AI).

7. Bibliographical References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate speech transcription of long-form audio. *Interspeech*, abs/2303.00747.

Feng-Ju Chang, Jing Liu, Martin H. Radfar, Athanasios Mouchtaris, et al. 2021. Context-aware transformer transducer for speech recognition. *ASRU*.

Jennifer Drexler Fox and Natalie Delworth. 2022. Improving contextual recognition of rare words with an alternate spelling prediction model. *Interspeech*.

Kaixun Huang, Aoting Zhang, Zhanheng Yang, Pengcheng Guo, et al. 2023. Contextualized end-to-end speech recognition with contextual phrase prediction network. *Interspeech*, abs/2305.12493.

Mahaveer Jain, Gil Keren, Jay Mahadeokar, and Yatharth Saraf. 2020. Contextual RNN-T for open domain ASR. In *Interspeech*.

Duc Le, Mahaveer Jain, Gil Keren, Suyoun Kim, et al. 2021. Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion. *Interspeech*, abs/2104.02194.

Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, and Louis-Philippe Morency. 2023. Multimodal lecture presentations dataset: Understanding multimodality in educational slides. *ICCV*.

Zhengyang Li, Thomas Graave, Jing Liu, Timo Lohrenz, et al. 2023. Parameter-efficient cross-language transfer learning for a language-modular audiovisual speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Da-Rong Liu, Chunxi Liu, Frank Zhang, Gabriel Synnaeve, et al. 2020. Contextualizing ASR lattice rescoring with hybrid pointer network language model. In *Interspeech*.

Edie Lu, Philip Harding, Kanthashree Mysore Sathyendra, Sibo Tong, et al. 2023. Model-internal slot-triggered biasing for domain expansion in neural transducer ASR models. In *Interspeech 2023*.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, et al. 2005. The AMI meeting corpus.

Salima Mdhaffar, Y. Estève, Antoine Laurent, Nicolas Hernandez, et al. 2020. A multimodal educational corpus of oral courses: Annotation, analysis and case study. In *LREC*.

Robert M Ochshorn and Max Hawkins. 2017. Gentle: A robust yet lenient forced aligner built on kaldi.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, et al. 2021. SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. In *Interspeech*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. *ICASSP*.

Rahul Pandey, Roger Ren, Qi Luo, Jing Liu, et al. 2023. Procter: Pronunciation-aware contextual adapter for personalized speech recognition in neural transducers. In *ICASSP*, pages 1–5.

Pradip Pramanick and Chayan Sarkar. 2022. Can visual context improve automatic speech recognition for an embodied agent? *EMNLP*, abs/2210.13189.

Golan Pundak, Tara N. Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao. 2018. Deep context: End-to-end contextual speech recognition. *SLT*.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *ACL*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, et al. 2022. Robust speech recognition via large-scale weak supervision. *ArXiv*, abs/2212.04356.

Swayambhu Nath Ray, Soumyajit Mitra, Raghavendra Bilgi, and Srinivas Garimella. 2021. Improving RNN-T ASR performance with date-time and location awareness. In *TDS*.

Miguel Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, et al. 2021. Earnings-21: A practical benchmark for ASR in the wild. *Interspeech*, abs/2104.11348.

Miguel Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra. 2022. Earnings-22: A practical benchmark for accents in the wild. *ArXiv*, abs/2203.15591.

Kanthashree Mysore Sathyendra, Thejaswi Muniyappa, Feng-Ju Chang, Jing Liu, et al. 2022. Contextual adapters for personalized speech recognition in neural transducers. In *ICASSP*.

Guangzhi Sun, C. Zhang, and Philip C. Woodland. 2022. Tree-constrained pointer generator with graph neural network encodings for contextual speech recognition. In *Interspeech*.

Haoxu Wang, Fan Yu, Xian Shi, Yuezhang Wang, et al. 2023a. SlideSpeech: A large-scale slide-enriched audio-visual corpus. *ArXiv*, abs/2309.05396.

Weiran Wang, Zelin Wu, Diamantino Caseiro, Tsendsuren Munkhdalai, et al. 2023b. Contextual biasing with the Knuth-Morris-Pratt matching algorithm. *ArXiv*, abs/2310.00178.

Xiaoqiang Wang, Yanqing Liu, Jinyu Li, Veljko Miljanic, et al. 2022. Towards contextual spelling correction for customization of end-to-end speech recognition systems. *IEEE/ACM TASLP*, 30.

Kai Wei, Thanh Tran, Feng-Ju Chang, Kanthashree Mysore Sathyendra, et al. 2021. Attentive contextual carryover for multi-turn end-to-end spoken language understanding. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 837–844.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, et al. 2023. BloombergGPT: A large language model for finance. *ArXiv*, abs/2303.17564.

Chao-Han Huck Yang, Linda Liu, Ankur Gandhe, Yile Gu, et al. 2021. Multi-task language modeling for improving speech recognition of rare words. *ASRU*.

Xiaoyu Yang, Wei Kang, Zengwei Yao, Yifan Yang, et al. 2023. PromptASR for contextualized ASR with controllable style. *ArXiv*, abs/2309.07414.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, et al. 2023. Zipformer: A faster and better encoder for automatic speech recognition.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. 2021. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45.

Ding Zhao, Tara N. Sainath, David Rybach, Pat Rondon, et al. 2019. Shallow-fusion end-to-end contextual biasing. In *Interspeech*.

A. Statistical Significance Testing

To assess whether the improvements in *ConEC* are statistically significant, we conducted paired t-test on all paired entity mentions between no biasing (row 1) and other rows in Table 2. The significance test results are shown in Table 2. For example, comparing row 1 (no biasing) and 4 (shallow fusion with ConEC) We observe that the WER improvements of PERSON, ORG, PRODUCT are statistically significant ($p < 0.05$), while GPE, FAC, LOC, EVENT, and NORP are not. This is not surprising since shallow fusion is not a very strong method to improve the no biasing baseline over all entity types. In particular, shallow fusion has difficulty with the types on which the no biasing has already low WERs. However, the oracle WER for ConEC (row 5), yields to statistically significant improvements compared to row 1 for all types, except for LOC, which indicates the potential of usefulness of the real context for contextual ASR. We believe there is significant room for future work to close the gap between the "no-biasing" and "oracle" results.

On the other hand, contrasting row 2 and row 4 shows that their differences are significant except for GPE, EVENT, and NORP. Contrasting row 3 and row 4 also shows significant differences for PERSON, PRODUCT, and ORG. This again suggests that real-world contexts (row 4) are noisier and more challenging than the synthesized contexts (rows 2 and 3).

B. Entity Types

We include a brief description of the entity types for self-containedness. Details of entity types can be found on https://github.com/revdotcom/speech-datasets/tree/main/earnings21

Table 4: Description of Entity Types

Entity Type	Description	Examples
PERSON	Names of people, including fictional people	Hagrid, Jason Chicola, W. E. B. Du Bois
ORG	Companies, agencies, institutions, etc.	Rev, General Motors, SEC, NAACP
GPE	Countries, cities, states, etc. Geopolitical entities.	Italy, US, Boston, New Zealand
LOC	Non-GPE locations, mountain ranges, bodies of water,etc.	the North, the Rocky Mountains
PRODUCT	Objects, vehicles, foods, etc. (not services)	Camry, Sufentanil, ARX-02
EVENT	Named hurricanes, battles, wars, sports events, etc.	COVID-19, the Spanish Flu, Hurricane Katrina, World War II
NORP	Nationalities or religious or political groups	American, Chinese, Republican, Grand Old Party, Roman Catholic
FAC	Buildings, airports, highways, bridges, etc.	Golden Gate Bridge, the Empire State Building