Earnings-21: A Practical Benchmark for ASR in the Wild

Miguel Del Rio\textsuperscript{1}, Natalie Delworth\textsuperscript{1}, Ryan Westerman\textsuperscript{1}, Michelle Huang\textsuperscript{1}, Nishchal Bhandari\textsuperscript{1}, Joseph Palakapilly\textsuperscript{1}, Quinten McNamara\textsuperscript{3}, Joshua Dong\textsuperscript{1}, Piotr Żelasko\textsuperscript{2,3}, Miguel Jette\textsuperscript{1}

\textsuperscript{1}Rev.com,
\textsuperscript{2}Center for Language and Speech Processing, \textsuperscript{3}Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD, USA

miguel.delrio@rev.com

Abstract

Commonly used speech corpora inadequately challenge academic and commercial ASR systems. In particular, speech corpora lack metadata needed for detailed analysis and WER measurement. In response, we present Earnings-21, a 39-hour corpus of earnings calls containing entity-dense speech from nine different financial sectors. This corpus is intended to benchmark ASR systems in the wild with special attention towards named entity recognition. We benchmark four commercial ASR models, two internal models built with open-source tools, and an open-source LibriSpeech model and discuss their differences in performance on Earnings-21. Using our recently released \texttt{fstalign} tool, we provide a candid analysis of each model’s recognition capabilities under different partitions. Our analysis finds that ASR accuracy for certain NER categories is poor, presenting a significant impediment to transcription comprehension and usage. Earnings-21 bridges academic and commercial ASR system evaluation and enables further research on entity modeling and WER on real world audio.

Index Terms: automatic speech recognition, named entity recognition, dataset, earnings call

1. Introduction

Automatic Speech Recognition (ASR) has been adopted for a wide variety of acoustic environments. Users expect ASR systems to understand a wide range of voices in various settings such as podcasts, quarterly earnings calls, and streaming video captioning. Whereas there exist multiple techniques that allow to adapt an ASR system to various acoustic conditions \cite{1,2,3}, it is also necessary to evaluate the system in its target operating conditions.

At present, common publicly available evaluation sets include LibriSpeech \cite{4}, Switchboard \cite{5}, CallHome \cite{6}, Rich Transcription 2007 \cite{7}, among others. The latest of these evaluation sets is over five years old, and none of them feature a wide variety of voices, technical domains, or acoustic environments. Recent corpora such as CHiME-5 \cite{8} poorly reflect real-world recording conditions. Furthermore, many of these traditional evaluation sets are not free to use, limiting access to research groups or well-funded private companies. The most challenging public test suite our team has used in the past is the AMI corpus \cite{9}, which features difficult speakers and good variance in recording characteristics. Recently, \cite{10} have shown that these standard ASR tasks and benchmarks create an overly-optimistic and misleading view of the current state of the art. Whereas the best reported word error rate (WER) results on LibriSpeech (around 1.4\% \cite{11}) or Switchboard (around 5\% \cite{12}) are encouraging, in reality, most commercially available systems are much closer to 15-20\% \cite{10} when transcribing user-provided recordings, making ASR a problem that is far from solved.

For transcription services such as Rev, the ASR API is domain agnostic, which necessitates a substantial effort in the procurement of evaluation sets that reflect a wide variety of acoustic environments, domains, voices, and accents.

In order to bolster the community’s efforts in robust ASR research, we release Earnings-21, an open and free evaluation corpus consisting of earnings call recordings and their corresponding rich transcripts available on Github\footnote{https://github.com/revdotcom/speech-datasets/tree/master/earnings21}. The main contributions of Earnings-21 are:

- A new freely available resource for ASR evaluation, sourced “in the wild” from recordings created during the year 2020
- Richly annotated transcripts (with punctuation, true-casing, and named entities) for detailed error analysis
- A benchmark of commercial and academic ASR systems on the corpus
- \texttt{fstalign} \footnote{https://github.com/revdotcom/fstalign}, a novel toolkit for quickly computing WER that leverages NER annotations

The rest of the paper is organized as follows: Section 2 details dataset properties and sourcing methodology, Section 3 compares the performance of various ASR systems on our new evaluation set, and Section 4 presents our future plans.

2. The Earnings-21 Dataset

The Earnings-21 dataset consists of 44 public earnings calls recorded in 2020 from 9 corporate sectors downloaded from Seeking Alpha\footnote{https://seekingalpha.com/earnings/earnings-call-transcripts}, totalling 39 hours and 15 minutes. Our data selection intends to reflect real world settings with diverse semantic and acoustic properties. The files in Earnings-21 contain:

- Varied sector-specific technical terminology
- A wide range of recording qualities - representative of audio typically received in the wild
- Entity-rich transcripts with annotated numerical figures
- Semantic and linguistic content unique to the year 2020

In particular, earnings calls have vastly varying recording characteristics and speaker profiles in the same call. We do not have...
any information on this audio metadata other than what can be inferred from the audios themselves. The audios are stored as monaural MP3 files.

To cover a wide range of scenarios common in real-world use cases, we chose recordings that had diverse sample rates as presented in Table 1. The recordings in this corpus range in length from less than 17 minutes to 1 hour and 34 minutes with the average recording being about 54 minutes in length.

| Sampling rate (Hz) | Recordings | Total time (hh:mm) |
|--------------------|------------|--------------------|
| 44100              | 7          | 07:13              |
| 24000              | 21         | 17:45              |
| 22050              | 5          | 04:12              |
| 16000              | 6          | 04:52              |
| 11025              | 5          | 05:14              |

Table 1: Sampling rate distribution across Earnings-21 in number of files and total duration.

2.1. Earnings call selection

Seeking Alpha defines 9 different sectors that categorize all earnings calls on their website - these are: Basic Materials, Conglomerates, Consumer Goods, Financial, Healthcare, Industrial Goods, Services, Technology, and Utilities. The average sector has just over 40,000 tokens and has about 4.5 hours of audio. To ensure diverse coverage, we randomly selected 5 calls that occurred in 2020 from each of these sectors.

2.2. Dataset transcription

To get accurate transcriptions, we used the Rev.com human transcription service. These audios were rigorously transcribed by a pool of experienced transcriptionists, then graded and reviewed by a different pool of senior transcriptionists. Spot-checking by the paper authors find that transcripts created using this process are highly accurate.

We chose to get “verbatim” transcripts that capture all speech utterances in exactly the same way those words were spoken – including filler words, false starts, grammatical errors, and other verbal cues or disfluencies. We found that verbatim transcripts are more useful for ASR evaluation. Real-world speech features frequent stuttering, repetition, and other disfluencies – modelling these mistakes is important for accurate transcription of a given recording [13][14].

During transcription, a transcriber found that one of the earnings calls contained a large amount of non-English speech; we removed this call from the dataset without replacing it because the remaining files still provide adequate coverage of entities over all sectors.

2.3. Data preparation

At Rev, we store reference transcripts in a custom format file we call .nlp files. These files are .csv-inspired pipe-separated (i.e. | ) text files that present tokens and their metadata on separate lines. We assigned NER labels to each transcript in three stages. First we used our internal NER tools to tag tokens that require text normalization such as abbreviations, cardinals, ordinals, and contractions. Next, we applied SpaCy 2.3.5’s NER tags to cover entities our labeller does not tag; these include organizations, people, and nationalities to name a few. Finally, we manually reviewed these tags and updated the entities. The labeled entities are distributed as shown in Figure 1.

As part of this release, we also include all metadata available to us. In some recordings, speakers are identified by name – when provided by the transcriptionists we include these in the speaker metadata. On a per-file basis, we take advantage of the metadata gathered as part of the data selection. This includes file length in seconds, file sampling rate, the company name / sector, the calls financial quarter, the number of unique speakers, and the total number of utterances in each recording.

3. Results on Earnings-21

We evaluated the transcription accuracy between four commercial ASR systems, two of our own ASR systems, and an open-source Kaldi model on Earnings-21. All models are run using an offline, batch decoding approach (the commercial models are run using their offline API pipeline when available). Using fstalign, we provide detailed WER analysis comparison of the earnings call transcription results.

3.1. WER calculation

As ASR systems become more accurate, more sophisticated measurement tools are needed to attenuate the effects of trivial, ambiguous, or otherwise less interesting errors. Our new open-source tool fstalign enables this by allowing for specific word substitutions and incorporating text normalization information.

To attenuate the effects of trivial errors, our tool uses a curated list of common word transforms that enable synonymous tokens to be leniently substituted. These transforms allow for hypothesis and reference transcripts to differ in semantically insignificant or ambiguous ways without penalizing WER scores. The following are two example transforms:

```
 going to → gonna
 I’ll → I will
```

In this example, for the reference “I’m going to win.”, “I’m gonna win.” would be a penalty-free hypothesis. In Earnings-21, these synonym-transformations typically affect 0.3% of the potential transcript disagreements.

4 This leaves the Conglomerates sector with only 4 calls.
5 We provide a detailed explanation of the nlp file format and a description of each entity class in our Github.
These providers are
surement of where industry stands with respect to our dataset.
the best in general so that we can get the best possible mea-
cial provider, we selected the model we’ve found to perform
the market to run our experiments against. For each commer-
3.2. Commercial models
We chose four of the best commercial cloud ASR providers in
the market to run our experiments against. For each commer-
cial provider, we selected the model we’ve found to perform
the best in general so that we can get the best possible mea-
urement of where industry stands with respect to our dataset.
These providers are Google (using the Video model), Amazon,
Microsoft, and Speechmatics. These models are all black-box
to us and therefore we cannot provide more information about
the specifications. The commercial model output is provided in
the data release for convenient reproducibility.

3.3. Internal models
We trained two models using the popular Kaldi and ESPNet
toolkits. Our models were developed as part of general ASR
systems with training data sourced from an unbiased selection
from our database. Audio is resampled at 16kHz for training
and inference time.

The first system is a Kaldi\cite{15} based DNN-HMM trained
on 30,000 hours of out-of-domain proprietary audio. The acous-
tic model is comprised of 80M parameters in interleaved TDNN
and LSTM layers\cite{16}, a 3-gram decoder (6M entries), and a 4-
gram LM (150M entries) interpolated with a 16M parameter
TDNN-LSTM RNNLM for rescoring.

The second system is an ESPNet V2\cite{17} based hybrid
CTC/Attention encoder-decoder\cite{18} model trained on 10,000
hours of out-of-domain proprietary audio. We used the prede-
fined LibriSpeech \texttt{conformer7} configuration with 124M param-
eters, 10,000 BPE\cite{19} tokens, and a maximum token length of
8 characters.

For further WER comparison, we also used an open-source
Kaldi model\cite{14} trained on 960 hours from LibriSpeech\cite{4}
with standard 3-way speed perturbation.

3.4. Comparison
We present the results of our WER measurements in Ta-
ble 2. We find that the ESPNet model is the most accurate
on Earnings-21. We were surprised to see the chasm between
the open-source LibriSpeech model and the proprietary ASR
systems. We posit that this is due to: (1) domain mismatch due
to vastly different acoustic channel and recording characteris-
tics between LibriSpeech and earnings calls, and (2) orders-of-
magnitude difference in amount of data used to train the ASR
models, as in the case of our models which used over 10 times
the amount of data used to train the open source models.

One recording\cite{7} showed significantly degraded WER on all
models. Manual review reveals that several speakers have heavy
accents and low recording quality; many models failed to tran-
scribe large sections of this file. If this file is excluded from
evaluation, our internal Kaldi model is the most accurate. We
have intentionally chosen to keep this difficult file as it presents
realistic lens into the variability of audio in the wild.

In the following subsections, we analyze the WER results
using different stratifications to better understand the nature of
ASR performance on real-world audio.

3.4.1. Entity recognition
We show how different models performed on the identified
named entity classes in Table 3. Looking for the entity classes
with the lowest WER across all models, we find \texttt{DATE}, \texttt{ORDI-
NAL}, and \texttt{TIME} are the easiest entity classes to recognize. We
hypothesize that these entities are easier due to their structured
pattern and frequent appearance in training data. On the other
hand, looking at the entities with the highest overall WER, we
find that \texttt{FAC}, \texttt{ORG}, and \texttt{PERSON} are difficult, which may be
\footnote{File-id 4346923 in the sector “Industrial Goods” and with sampling
rate of 16kHz}
Table 4: WER breakdown by sector (domain) defined by Seeking Alpha. We denote a model’s best and worst performing sector with ▽ and + respectively.

| Sector            | Google | Amazon | Microsoft | Speechmatics | Rev | ESPNet | LibriSpeech |
|-------------------|--------|--------|-----------|--------------|-----|--------|-------------|
| **Mean Sector**   | 17.8   | 17.1   | 15.8      | 16.0         | 13.2 | 11.3   | 48.8        |
| Conglomerates     | 15.5▽ | 15.4▽ | 14.1      | 14.0▽        | 8.0  | 10.2   | 44.1        |
| Utilities         | 15.9   | 15.9   | 14.8      | 14.2         | 10.3 | 10.8   | 45.7        |
| Basic Materials   | 16.7   | 15.5   | 14.6      | 14.5         | 11.0 | 11.1   | 43.6        |
| Services          | 16.8   | 16.6   | 14.8      | 15.2         | 11.5 | 9.8▽  | 44.1        |
| Healthcare        | 17.1   | 17.1   | 15.6      | 16.0         | 11.0 | 10.6   | 44.6        |
| Financial         | 18.0   | 17.0   | 15.6      | 15.5         | 13.2 | 11.5   | 49.5        |
| Consumer Goods    | 18.7   | 17.3   | 16.0      | 16.1         | 12.1 | 10.3   | 51.1        |
| Technology        | 20.6   | 18.9   | 17.1      | 17.4         | 16.0 | 12.9   | 56.3        |
| Industrial Goods  | 21.2+  | 20.0▽ | 19.3▽     | 21.0▽        | 25.9+| 14.4+  | 60.2+       |

Table 5: WER breakdown by recording’s sampling rate. We denote a model’s best and worst performing sample rate with ▽ and + respectively. See Table 4 for information on the distribution of sample rates.

| Sample Rate (Hz) | Google | Amazon | Microsoft | Speechmatics | Rev | ESPNet | LibriSpeech |
|------------------|--------|--------|-----------|--------------|-----|--------|-------------|
| **Mean Sample Rate** | 18.1   | 17.5   | 16.2      | 16.4         | 14.5 | 11.8   | 49.0        |
| 44100            | 16.0   | 15.5▽  | 14.9      | 14.4         | 10.0 | 9.6▽  | 40.3▽       |
| 24000            | 17.3   | 16.3   | 15.0      | 15.2         | 11.3 | 10.4   | 49.7        |
| 22050            | 14.6▽ | 15.6   | 13.4▽     | 12.6▽        | 8.9▽| 10.5   | 43.3        |
| 16000            | 22.0+  | 21.1▽  | 20.4+     | 22.3+        | 28.1+| 15.1+  | 59.5+       |
| 11025            | 19.9   | 19.1   | 17.2      | 17.5         | 14.2 | 13.5   | 52.2        |

due to higher lexical diversity in those categories. ASR systems incorporate language models which are particularly sensitive to sparsity, making recognition of rare or novel token sequences difficult without special modeling.

3.4.2. Domain

We demonstrate the WER results with topic-domain partitioning as defined by business sector in Table 4. The data set shows poorest accuracy in Industrial Goods and Technology sectors and best accuracy in the Conglomerates sector. The measured difficulty of transcribing Industrial Goods is attributed to the most difficult file in the corpus; treating that file as an outlier leads us to believe the sector’s difficulty is average. In the Technology domain, models suffer from large contiguous deletions which can account for over 40% of the errors in a file. More work needs to be done to understand the discrepancies in accuracy between sectors.

3.4.3. Sampling rate

The data set recordings have diverse sampling rates. We compare ASR performance with respect to sampling rate in Table 5. We find that most systems perform similarly at 22050Hz and greater sampling rates. We note that our most difficult file to transcribe is 16kHz and skews the WER averages, but omitting this file shows a clear positive correlation between sampling rate and accuracy. 11025Hz is a particularly difficult sampling rate, which can account for over 40% of the errors in a file. More work needs to be done to understand the discrepancies in accuracy between sectors.

4. Conclusion

We show that there still exist major obstacles to speech recognition in the wild. With our data set release, we challenge researchers to deal with real-world audio. We also provide fstalign as a tool to enable the research community to focus on higher-importance entity WER and move past trivial errors.

We will continue to improve the metadata for our Earnings-2I corpus and invite others to contribute as well. We hope this release is the first of many towards providing a realistic view of speech in the wild. We encourage industry leaders and academic researchers to continue research in this vein, as continued efforts towards modeling real-world challenges and up-to-date data will be the future of ASR.

5. Acknowledgements

We would like to thank the transcriptionists and contractors who dedicated hundreds of hours and helped us to create and improve the quality of this dataset.

6. References

[1] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” arXiv preprint arXiv:1510.08484, 2015.
[2] X. Cui, V. Goel, and B. Kingsbury, “Data augmentation for deep neural network acoustic modeling,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 9, pp. 1469–1477, 2015.
[3] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “SpeechAugment: A simple data augmentation method for automatic speech recognition,” arXiv preprint arXiv:1904.08779, 2019.
[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015, pp. 5206–5210.
