SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition

Patrick K. O’Neill1, Vitaly Lavrukhin2, Somshubra Majumdar2, Vahid Noroozi2, Yuekai Zhang3, Oleksii Kuchaiev2, Jagadeesh Balam2, Yuliya Dovzhenko1, Keenan Freyberg1, Michael D. Shulman1, Boris Ginsburg2, Shinji Watanabe3,4, Georg Kucsko1

1Kensoh Technologies, Cambridge MA, USA
2NVIDIA Technologies, Santa Clara CA, USA
3Johns Hopkins University, Baltimore MD, USA
4Carnegie Mellon University, Pittsburgh PA, USA

patrick.oneill@kensho.com, georg@kensho.com

1We concede that the precise details of English orthography can vary by time, geographical region, and even publication house style. We nevertheless ostensively define orthographic text here to mean “text as it generally appears in U.S. print publications”.

2Rhymes with “squeegee”.

1. Abstract

In the English speech-to-text (STT) machine learning task, acoustic models are conventionally trained on uncased Latin characters, and any necessary orthography (such as capitalization, punctuation, and denormalization of non-standard words) is imputed by separate post-processing models. This adds complexity and limits performance, as many formatting tasks benefit from semantic information present in the acoustic signal but absent in transcription. Here we propose a new STT task: end-to-end neural transcription with fully formatted text for target labels. We present baseline Conformer-based models trained on a corpus of 5,000 hours of professionally transcribed earnings calls, achieving a CER of 1.7. As a contribution to the STT research community, we release the corpus free for non-commercial use.

2. Introduction

In the English speech-to-text (STT) task, acoustic model output typically lacks orthography, the full set of conventions for expressing the English language in writing (and especially print). Acoustic models usually render uncased Latin characters, and standard features of English text such as capitalization, punctuation, denormalization of non-standard words, and other formatting information are omitted. While such output is suitable for certain purposes like closed captioning, it falls short of the standard of English orthography expected by readers and can even pose problems for certain downstream NLP tasks such as natural language understanding, neural machine translation and summarization. In contexts where standard English orthography is required, therefore, it is typically provided after the fact by a pipeline of post-processing models that infer a single aspect of formatting each from the acoustic model output.

This approach suffers several drawbacks. First, rendering orthographically correct text is the more natural task: if asked to transcribe audio and given no further instructions, a normally literate English speaker will tend to generate orthographic text. By contrast, writing in block capitals without punctuation, spelling out numbers in English and so on, is a far less conventional style. Secondly, certain types of orthographic judgments are possible only with acoustic information and cannot be reliably inferred from text alone. Consider, for example, the problem of inferring the correct EOS marker for the sentence “the CEO retired” (either of a period or question mark) without the information carried in vocal pitch. Third, the practice of chaining orthography-imputing models into pipelines tends to encumber STT systems. Each orthographic feature may require a model of significant engineering effort in its own right (cf. [4, 5, 6, 7, 8, 9]), and improvements to one model’s performance may degrade end-to-end performance as a whole due to distribution shift.

Modern STT models require large volumes of high-quality training data, and to our knowledge there is no extant public corpus suitable for the fully-formatted, end-to-end STT task. To address this limitation we release SPGISpeech, a subset of the financial audio corpus described in [11]. SPGISpeech offers:

- 5,000 hours of audio from real corporate presentations
- Fully-formatted, professional manual transcription
- Both spontaneous and narrated speech
- Varied teleconference recording conditions
- Approximately 50,000 speakers with a diverse selection of L1 and L2 accents
- Varied topics relevant to business and finance, including a broad range of named entities.

3. Prior Work

There are many extant STT corpora, varying in their volumes as well as in the details of their formats. Early corpora include the Wall Street Journal corpus, consisting of 80 hours of narrated news articles [12]. SWITCHBOARD, containing approximately 300 hours of telephone conversations [13]; TIMIT, consisting of a set of ten phonetically balanced sentences read by hundreds of speakers (~ 50 h) [14]; and the Fisher corpus of transcribed telephone conversations (2,000 h) [15].

The LibriSpeech corpus, a standard benchmark for STT models, consists of approximately 1000 hours of narrated audiobooks [16]. Although the use of audiobooks as an STT corpus allows researchers to leverage a large pre-existing body of transcription work, this approach poses several limitations. First, LibriSpeech consists entirely of narrated text, hence it...
lacks many of the acoustic and prosodic features of spontaneous speech. Second, as the narrated books are public domain texts, there is a bias in the corpus towards older works, and hence against more modern registers of English.

Other corpora include TED-LIUM (450 hours of transcribed TED talks) [17] and Common Voice (a multilingual corpus of narrated prompts with ~1,100 validated hours in English) [18], and GigaSpeech (10,000 hours from audiobooks, podcasts and YouTube videos) [19]. We present a comparison of SPGISpeech to peer corpora in Table 1. We excluded all slices that contain currency information (8% of total), on the grounds that currency utterances often have non-trivial denormalizations and misquotation issues that require global context in order to render correctly. The utterance “‘one twenty three’“, for example, might be correctly transcribed as any of $1.23, £1.23, \$1.23\text{ million}$, and so on, depending on context. Given the potential for material errors in a business setting if reported incorrectly, moreover, misquotation of currency values in spontaneous speech are typically corrected in transcription. Lacking the means to verify the correct spoken form for each currency mention, we simply exclude them.

1. We sampled no more than four consecutive slices from any call. We also redacted the corpus out of concern for individual privacy. Though earnings calls are public, we nevertheless identified full names with the spaCy `en_core_web_large` model [23], which we selected on grounds of its wall clock performance for scanning the entire corpus. We withheld slices containing names that appeared fewer than ten times (7% of total). Full names appearing ten times or more in the data were considered to be public figures and were retained. This necessarily incomplete approach to named entity recognition was complemented with randomized manual spot checks which uncovered no false negatives missed by the automated approach.

2. We excluded all slices that contain currency information (8% of total), on the grounds that currency utterances often have non-trivial denormalizations and misquotation issues that require global context in order to render correctly. The utterance “‘one twenty three’“, for example, might be correctly transcribed as any of $1.23, £1.23, \$1.23\text{ million}$, and so on, depending on context. Given the potential for material errors in a business setting if reported incorrectly, moreover, misquotation of currency values in spontaneous speech are typically corrected in transcription. Lacking the means to verify the correct spoken form for each currency mention, we simply exclude them.

3. We excluded slices with transcriptions containing non-ASCII characters. In particular we excluded all slices whose transcripts did not consist entirely of the upper- and lowercase ASCII alphabet; digits; the comma, period, apostrophe, hyphen, question mark, percent sign, and space characters.

4. Remaining slices were randomly subsampled in order to construct the published corpus, which consists of two splits, `train` and `val`, having no call events in common. We also construct a private `test` split, defined exactly as `val` and having no calls in common with it, which we do not release.

4. Corpus Definition

The SPGISpeech corpus is derived from company earnings calls manually transcribed by S&P Global, Inc. according to a professional style guide detailing conventions for capitalization, punctuation, denormalization of non-standard words and transcription of disfluencies in spontaneous speech. The basic unit of SPGISpeech is a pair consisting of a 5 to 15 second long 16 bit, 16kHz mono `wav` audio file and its transcription.

4.1. Alignment and Slicing

Earnings calls last 30-60 minutes in length and are typically transcribed as whole units, without internal timestamps. In order to produce short audio slices suitable for STT training, the files were segmented with Gentle [21], a double-pass forced aligner, with the beginning and end of each slice of audio imputed by voice activity detection with `py-webrtc` [22]. While there is inevitably a potential to introduce certain systematic biases through this process, the fraction of recovered aligned audio per call ranges from approximately 40% in the calls of lowest audio quality to approximately 70% in the highest.

4.2. Corpus Definition

Slices in SPGISpeech are not a simple random sample of the available data, but are subject to certain exclusion criteria.
5. Corpus Analysis

We briefly characterize the data. Summary statistics are given in Table 3. Several representative examples are given in Table 2 highlighting some of the challenges SPGISpeech presents for end-to-end training. The table contains samples from the validation split where the correct EOS marker is difficult to infer from the transcription of the slice alone, non-standard words that must be denormalized, disfluencies and hesitations arising from self-correction in spontaneous speech, and instances like year abbreviations where semantic information is likely necessary to determine the correct orthography. In each instance we also report characteristic Conformer model output (see Section 6), demonstrating the feasibility of end-to-end orthographic transcription for these examples.

SPGISpeech contains many specialized forms and entity types such as acronyms (15% of all slices), pauses (10%), organizations (25%), persons (8%), and locations (8%). For each form we estimate its prevalence from a random sample of transcripts. Prevalences of the named entity types person, organization and location were estimated using Flair [24], which we selected on grounds of its accuracy and acceptable wall clock performance for scanning a small sample of the corpus.

There are roughly 50,000 speakers in SPGISpeech, drawn from corporate officers and analysts appearing in English earnings calls and spanning a broad cross-section of L1 and L2 accents. We attempted to characterize this linguistic diversity by tabulating the domiciles of the corporate headquarters of the companies in the corpus. In Fig. 1 we report the distribution of the global region associated with each company’s headquarters as listed in the S&P Capital IQ database [25]. To further refine the picture of accent composition, we next considered the distribution of countries of corporate domicile. We found it to be long-tailed, the top five most common nations (US, Japan, UK, India, Canada) comprising 2/3 of the data. Although a few of the twenty most frequent countries are reputed for encouraging strategic corporate domiciling, we find that their combined share of the total is low, suggesting that their inclusion does not significantly distort the bulk statistics. While imputation of speaker accent from corporate domicile is a necessarily limited approach, we found the observed statistics to be broadly concordant with direct estimation of accent in a random sample. The number of speakers and diversity of accents in SPGISpeech is, to our knowledge, the widest of any public corpus of spontaneous speech, making it especially suitable for training robust STT models.

Lastly, we analyzed the industrial composition of included companies in order to ensure that they constituted a representative cross-section of business topics. All eleven top-level sectors of the Global Industry Classification Standard (GICS) [26] are reflected in the data at rates roughly comparable to their proportions in the US economy. SPGISpeech, not being constrained

Table 2: Examples of non-standard text in SPGISpeech. For each textual feature we give an example of the ground truth transcript label, a verbatim transcription without casing or punctuation, and Conformer model output with end-to-end orthographic training.

Table 3: SPGISpeech Summary Statistics.

![Distribution of Speakers by Global Region](image)

Figure 1: Distribution of Speakers by Global Region. Speaker distribution is estimated in a random sample according to reported region of corporate domicile.
to any particular company or industrial sector, therefore offers a fairly synoptic view of modern English business discourse.

6. Transcription Experiments

To illustrate the feasibility of the end-to-end orthographic approach we train several models on SPGISpeech, including Conformer (ESPnet) and Conformer-CTC (Nemo).

6.1. Model Descriptions and Methods

Both presented models are based on the recently proposed Conformer (Convolution-Augmented Transformer) architecture \[27\] which combines the local sensitivities of convolutional neural networks \[28, 29\], with the long-range interactions of transformers \[30\] in order to capture both local and global dependencies in audio sequences. The unit of prediction for both models consists of SentencePiece tokenized text \[31\], whose detokenized output is upper- and lower-case characters, digits, the comma, period, apostrophe, hyphen, question mark, percent sign, and the space character, covering the full range of characters present in SPGISpeech.

The ESPNet Conformer model presented consists of 12 conformer blocks with an output dimension of 512 and a kernel size of 31 in the encoder, and 6 transformer blocks in the decoder. Both encoder and decoder have 8 attention heads with 2048 feed-forward unit dimension. The size of the vocabulary was chosen at \(\sim 5k\). The model was trained using four 24Gb memory Titan RTX GPUs for 35 epochs. Adam optimizer with no weight decay was used and Noam learning rate scheduler was applied with 25k warmup steps and a learning rate of 0.0015. SpecAug was used with 2 frequency masks and 5 time masks. The last 10 best checkpoints were averaged for the final model. Detailed setups can be found in the ESPNet recipe.

The Conformer-CTC model, provided through the NeMo toolkit \[32\], is a CTC-based variant of the Conformer which has the same encoder but uses CTC loss \[33\] instead of RNNT \[34\]. It also replaces the LSTM decoder with a linear decoder on the top of the encoder. This makes Conformer-CTC a non-autoregressive model unlike the original Conformer, allowing for significantly faster inference speeds. The presented model was trained for 230 epochs at an effective batch size of 4096, with Adam optimizer and no weight decay. SpecAug was applied with 2 frequency maskings of 27, and 5 time maskings at a maximum ratio of 0.05. Noam learning scheduler was used with a warmup of 10k steps and learning rate of 2.0.

6.2. Results and Discussion

As performance on the orthographic STT task hinges in large part on single-character distinctions such as capitalization and punctuation, we report both WER and CER for the test split of SPGISpeech, closely tracking the performance on the val split. We also estimate the respective error rates for a normalized transcription task by lowering the output and filtering all but letters, the apostrophe and the space character to conform to the conventional choice of STT vocabulary. Both results are shown in Table 4.

Our results demonstrate CERs in the orthographic STT task comparable to those obtained on standard normalized corpora. While we caution against laying undue stress upon any one particular comparison made across studies, they suggest on the whole that English orthography is within the grasp of modern acoustic architectures, and hence end-to-end orthographic STT is a feasible task. Both models achieve WERs of less than 6.0 and CERs less than 2.0, making them broadly comparable to previous results obtained with Conformer in other corpora \[27\].

A comparison of the ortho and norm error rates suggests that 1/2 to 2/3rd of the error is due to orthographic issues. One must recall, however, that there is a lower bound imposed by irreducible error: in some cases there is genuine disagreement as to whether a pause merits a comma, or whether a hesitation should be recorded or emended. For the disfluency given in Table 4, for example, the model smoothly emends the first repeated word but records the second correction while omitting the partially pronounced word “limit”: some degree of judgment in the training labels, and hence irreducible error, is inevitable.

7. Conclusions

In this work we introduced a new end-to-end task of fully formatted speech recognition, in which the acoustic model learns to predict complete English orthography. This approach is both conceptually simpler and can lead to improved accuracy due to acoustic cues only present in the the audio, which are needed for full formatting. We demonstrated the feasibility of this approach by training models on SPGISpeech, a corpus uniquely suited for the task of large scale, fully formatted end-to-end transcription in English. As a contribution to the STT research community, we also offer SPGISpeech for free academic use.

8. Acknowledgements

The authors wish to thank the S&P Global Market Intelligence Transcripts team for data collection and annotation; Bhavesh Dayalji, Abhishek Tomar, Richard Neale and Gabriela Pereyra for their project support; and Shea Hudson Kerr and Stacey Steele for helpful legal guidance.

9. References

[1] K. H. Albrow, “The English writing system: Notes towards a description,” Schools Council Program in Linguistics and English Teaching, papers series 2, no. 2, 1972.

[2] A. Ravichander, S. Dalmia, M. Ryskina, F. Metze, E. Hovy, and A. W. Black, “NoiseQA: Challenge Set Evaluation for User-Centric Question Answering,” in Conference of the European Chapter of the Association for Computational Linguistics (EACL), Online, April 2021. [Online]. Available: https://arxiv.org/abs/2102.08345
