INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

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

Patrick K. O'Neill[1], *Vitaly Lavrukhin*[2], *Somshubra Majumdar*[2], *Vahid Noroozi*[2], *Yuekai Zhang*[3], *Oleksii Kuchaiev*[2], *Jagadeesh Balam*[2], *Yuliya Dovzhenko*[1], *Keenan Freyberg*[1], *Michael D. Shulman*[1], *Boris Ginsburg*[2], *Shinji Watanabe*[3,4], *Georg Kucsko*[1]

[1]Kensho Technologies, Cambridge MA, USA
[2]NVIDIA Technologies, Santa Clara CA, USA
[3]Johns Hopkins University, Baltimore MD, USA
[4]Carnegie Mellon University, Pittsburgh PA, USA
patrick.oneill@kensho.com, georg@kensho.com

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.[1]

1. 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)[2] [1]. 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 [2, 3]. 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 [10].

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[3], 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.

2. 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 ($\sim$ 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

[1]`https://datasets.kensho.com/datasets/scribe`

[2]We 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".

[3]Rhymes with "squeegee".

http://dx.doi.org/10.21437/Interspeech.2021-1860

Table 1: *Comparison of SPGISpeech to peer corpora. For a select group of comparable STT corpora, we compare the recording format, discursive domain, metadata and quantity of audio. Numeric data are rounded; precise values for SPGISpeech are given in Table 3.*

Corpus name	Acoustic condition	Speaking style
Switchboard	Telephone	Spontaneous
Librispeech	Close-talk mic.	Narrated
TedLium-3	Close-talk mic.	Narrated
Common Voice	Teleconference	Narrated
SPGISpeech	Teleconference	Spontaneous, Narrated

Corpus name	Transcription Style	Speaker Count	Vocabulary Size	Amount (h)
Switchboard	Orthographic	550	25,000	300
Librispeech	Non-orthographic	2,400	200,000	960
TedLium-3	Non-orthographic	2,000	160,000	450
Common Voice	Non-orthographic	40,000	220,000	1,400
SPGISpeech	Orthographic	50,000	100,000	5,000

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 $\sim$ 1,100 validated hours in English) [18], and GigaSpeech (10,000 hours from audiobooks, podcasts and YouTube videos) [19]. We present a comparison to select corpora in Table 1; a more exhaustive catalogue of prior STT corpora can be found in [20].

Within this field of previous work, SPGISpeech is distinctive for being over ten times larger than the next largest corpus with orthographic ground truth labels. It also contains approximately 50,000 speakers, the largest number to our knowledge of any public corpus.

3. 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 $\sim$10 second long 16 bit, 16kHz mono `wav` audio file and its transcription.

3.1. 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. Slice length was constrained to between 5 and 15 seconds in order to render audio in a convenient form for training and maximize comparability to extant datasets. Approximately half of all slices nevertheless contain an internal sentence boundary.

3.2. Corpus Definition

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

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 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.

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.*

Punctuation	
Transcript:	early in April versus what was going on at the beginning of the quarter?
Verbatim:	early you know in april versus uh what was going on at the beginning of the quarter
Model Output:	early in April versus what was going on at the beginning of the quarter?
Non-standard words	
Transcript:	[...] for the first time in our 92-year history, we [...]
Verbatim:	[...] for the first time in our ninety two year history we [...]
Model Output:	[...] for the first time in our 92-year history, we [...]
Disfluency	
Transcript:	As respects our use of insurance to put out -- reinsurance to put out [...]
Verbatim:	as as respects our use of insurance to put out lim reinsurance to put out [...]
Model:	As respects our use of insurance to put out -- reinsurance to put out [...]
Abbreviations	
Transcript:	in '15, and got margins back to that kind of mid-teens level [...]
Verbatim:	in fifteen and got margins back to that kind of mid teens level [...]
Model Output:	in '15 and got margins back to kind of that mid-teens level [...]

Table 3: *SPGISpeech Summary Statistics.*

	Train	Val
Events	55,289	1,114
Slices	1,966,109	39,341
Time (h)	5,000	100
Vocabulary	100,166	19,865
OOV	—	703

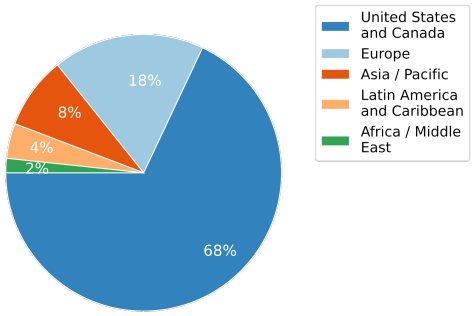

Figure 1: *Distribution of Speakers by Global Region. Speaker distribution is estimated in a random sample according to reported region of corporate domicile.*

4. 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 5), 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. All speakers are adults, and an estimated 80% are male. Speakers span a broad cross-section of L1 and L2 accents. We characterized diversity of accent 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 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, with approximately half of all nations represented yet the five most common (US, Canada, UK, India, Germany) comprising 3/4$^{\text{ths}}$ of the data. 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 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 in rough proportion to the broader US economy. SPGISpeech, not being constrained to any particular company or industrial sector, therefore offers a fairly synoptic view of modern English business discourse. Even for unrelated target domains, moreover, previous work has found financial audio to be suitable for cross-domain adaptation [11].

5. 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).

5.1. 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 ∼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[4].

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.

5.2. 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 lowercasing the output and filtering all but letters, the apostrophe and the space character to conform to the conventional choice of STT vocabulary. We lastly report the mean F1 score for each character class of punctuation marks, digits, uppercase and lowercase letters. F1 scores were calculated under global alignment of the predicted and true transcript pairs with scoring parameters implied by the definition of CER [35]. 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 2/3^{rds} 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. Nevertheless, the distribution of class-wise F1 scores suggests that punctuation and casing are challenging but learnable in the neural setting.

Table 4: *Model Results. Word error rate and character error rate for greedy decoding on the SPGISpeech* `test` *set, along with mean F1 scores for various character classes.* `ortho` *numbers refer to the unnormalized, fully formatted orthographic output, while* `norm` *numbers refer to a lowercased and reduced vocabulary to remove the effects of text formatting.*

		Conformer (ESPnet)	Conformer-CTC (NeMo)
WER	ortho	5.7	6.0
	norm	2.3	2.6
CER	ortho	1.7	1.8
	norm	1.7	1.8
F1 score	punctuation	0.86	0.86
	digits	0.99	0.99
	uppercase	0.89	0.90
	lowercase	0.99	0.99

6. 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.

7. 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.

8. 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

[4] `https://github.com/espnet/espnet/tree/master/egs2/spgispeech`

[3] S. Peitz, M. Freitag, A. Mauser, and H. Ney, "Modeling punctuation prediction as machine translation," in *in Proceedings of the International Workshop on Spoken Language Translation (IWSLT*, 2011.

[4] R. Sproat and N. Jaitly, "RNN approaches to text normalization: A challenge," *ArXiv*, vol. abs/1611.00068, 2016.

[5] W. Salloum, G. Finley, E. Edwards, M. Miller, and D. Suendermann-Oeft, "Deep learning for punctuation restoration in medical reports," in *BioNLP 2017*. Vancouver, Canada,: Association for Computational Linguistics, Aug. 2017, pp. 159–164. [Online]. Available: https://www.aclweb.org/anthology/W17-2319

[6] E. Pusateri, B. R. Ambati, E. Brooks, O. Plátek, D. McAllaster, and V. Nagesha, "A mostly data-driven approach to inverse text normalization," in *INTERSPEECH*, 2017.

[7] B. Nguyen, V. B. H. Nguyen, H. Nguyen, P. N. Phuong, T. Nguyen, Q. T. Do, and L. C. Mai, "Fast and accurate capitalization and punctuation for automatic speech recognition using transformer and chunk merging," in *22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques, O-COCOSDA 2019, Cebu, Philippines, October 25-27, 2019*. IEEE, 2019, pp. 1–5. [Online]. Available: https://doi.org/10.1109/O-COCOSDA46868.2019.9041202

[8] O. Hrinchuk, M. Popova, and B. Ginsburg, "Correction of automatic speech recognition with transformer sequence-to-sequence model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 05 2020, pp. 7074–7078.

[9] M. Sunkara, C. Shivade, S. Bodapati, and K. Kirchhoff, "Neural inverse text normalization," 2021.

[10] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

[11] J. Huang, O. Kuchaiev, P. O'Neill, V. Lavrukhin, J. Li, A. Flores, G. Kucsko, and B. Ginsburg, "Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition," in *International Conference on Multimedia and Expo*, 2021, to appear.

[12] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[13] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *ICASSP*, 1992.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," 1993.

[15] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a resource for the next generations of speech-to-text," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2004/pdf/767.pdf

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[17] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation," 2018.

[18] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," 2019.

[19] G. Chen, S. Chai, G. Wang, J. Du, C. W. Wei-Qiang Zhang, D. Su, D. Povey, J. Trmal, J. Zhang, M. Ji, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio," *submitted to INTERSPEECH*, 2021.

[20] J. Le Roux and E. Vincent, "A categorization of robust speech processing datasets," Mitsubishi Electric Research Labs TR2014-116, Technical Report, 2014.

[21] lowerquality, *Gentle Aligner*, 2020 (accessed May 4th, 2020). [Online]. Available: https://lowerquality.com/gentle/

[22] J. Wiseman, *pyWebRTC*, 2016 (accessed May 4th, 2020). [Online]. Available: https://github.com/wiseman/py-webrtcvad

[23] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[24] A. Akbik, D. Blythe, and R. Vollgraf, "Contextual string embeddings for sequence labeling," in *COLING 2018, 27th International Conference on Computational Linguistics*, 2018, pp. 1638–1649.

[25] S. Global, "S&p capital iq," www.capitaliq.com, 2012, accessed June 2020.

[26] M. S. C. International, "The global industry classification standard (GICS)," 2019. [Online]. Available: https://www.msci.com/gics

[27] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[28] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, pp. 193–202, 1980.

[29] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, 2014. [Online]. Available: https://doi.org/10.1109/TASLP.2014.2339736

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[31] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. [Online]. Available: https://www.aclweb.org/anthology/D18-2012

[32] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook, P. Castonguay, M. Popova, J. Huang, and J. M. Cohen, "Nemo: a toolkit for building ai applications using neural modules," 2019.

[33] A. Graves, S. Fernández, and F. Gomez, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *In Proceedings of the International Conference on Machine Learning, ICML 2006*, 2006, pp. 369–376.

[34] A. Graves, "Sequence transduction with recurrent neural networks," in *ICML 29*, 2012.

[35] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.*, vol. 48, pp. 443–453, 1970.