INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Human Transcription Quality Improvement

Jian Gao[1], *Hanbo Sun*[1], *Cheng Cao*[1], *Zheng Du*[1]

[1]Amazon, USA

gajian@amazon.com, sunhanbo@amazon.com, chengcao@amazon.com, zhengdu@amazon.com

Abstract

High quality transcription data is crucial for training automatic speech recognition (ASR) systems. However, the existing industry-level data collection pipelines are expensive to researchers, while the quality of crowdsourced transcription is low. In this paper, we propose a reliable method to collect speech transcriptions. We introduce two mechanisms to improve transcription quality: confidence estimation based reprocessing at labeling stage, and automatic word error correction at post-labeling stage. We collect and release LibriCrowd - a large-scale crowdsourced dataset of audio transcriptions on 100 hours of English speech. Experiment shows the Transcription WER is reduced by over 50%. We further investigate the impact of transcription error on ASR model performance and found a strong correlation. The transcription quality improvement provides over 10% relative WER reduction for ASR models. We release the dataset and code to benefit the research community.

Index Terms: speech transcription, crowdsourcing, dataset, confidence estimation, error correction

1. Introduction

Speech recognition systems have made significant progress during the recent years. The research community is actively developing algorithms for automatic speech recognition (ASR), and many models are at par with or even surpass humans [1]. However, less work has been published on the technology of manual speech recognition (MSR), though high quality human speech transcription is crucial for all spoken language related research.

The development of speech corpus produced by MSR follows the design of ASR system. Two decades ago, hybrid systems [2] were popular, in which Hidden Markov Models (HMMs) learn the transition between phones, and Recurrent Neural Networks (RNNs) perform localized classifications. Speech corpus like TIMIT [3] provides manually time-aligned phonetic and word transcriptions to support phone-level training and evaluation. Due to the high cost of hand-aligning each word with its corresponding audio frames, people use forced alignment [4] to automatically generate word aligned datasets. Later, CTC [5] was proposed to train ASR models on unsegmented data by aggregating over all possible alignments, and word-aligned datasets are no longer necessary.

In the 2010s, ASR systems were greatly improved and became hungry for data. TIMIT was found too small to train the prediction network in RNN-T [6]. This trend is similar as in the vision and language domains. GPT-3 [7] shows language model (LM) performance scales as a power-law of dataset size, model size, and computation power. Compared with text and image datasets, the scale of speech corpus is lagging far behind and many tasks are short of supervised data. To solve this problem, many recent works [8, 9, 10] move to unsupervised or self-supervised learning that use mostly unlabeled audio plus a small amount of transcribed speech. For example, with 10-hour labeled speech , WavLM [11] achieves similar word error rate (WER) as many fully supervised models [12, 13, 14] trained on the entire 960-hour LibriSpeech corpus [15]. Nevertheless, the data problem cannot be bypassed as labeled data is crucial for real-world applications in which the ASR model performance is often much worse than published benchmark results. To bridge this gap, the investment on quality controlled data collection is as important as improving ASR algorithms.

Today's industry-level data annotation pipelines [16] are often expensive and inconvenient for researchers, even though they hire professional transcribers. The standard transcription cost in U.S. is around $90 per speech hour. To save cost, some data providers use synthetic data or ASR-assisted pre-labeling to replace humans, which can introduce unwanted bias [17].

Crowdsourcing has been prevalent to produce the majority of labeled data in many computer vision tasks, replacing professional annotators. Platforms like Microsoft UHRS, Amazon MTurk, Appen, and Scale AI distribute annotation tasks to a large group of people, which is often cheap and efficient. However, for complex tasks like speech recognition, crowdsourced annotators who lack certified training are likely to provide low-quality annotations.

Thus far, few research has quantitatively analyzed the impact of human transcription error on ASR models. [18] claims HMM-based ASR systems are robust to mislabeled transcriptions as Gaussian Mixture Models learn very little from incorrect labels. However, [19] has conflicting observations on RNN-T based ASR systems. It shows the negative impact brought by transcription defects, especially the deletion error, cannot be recovered despite increasing model size and data size. Nevertheless, the transcription errors found in both [18] and [19] are from the simulation rather than by real human transcribers. [19] uses soundex [20] and bi-gram language model to simulate substitution and insertion errors. The simulated error patterns cannot well represent the mistakes made by real human.

In this paper, we propose a ML-in-the-loop data collection mechanism for high-quality speech transcription. It includes two schemes of quality improvement. First, during annotator's labeling, confidence estimation modules (CEMs) identify low-quality transcriptions for relabeling. Next, after annotator's labeling, error correction models (ECMs) generate the final output with reduced Transcription WER (TWER). Through experiments we find the two stages together provide a 50% TWER reduction. To analyze the transcription error's impact on training ASR models, we train Wav2Vec2 [8] and WavLM [11] on noisy transcription labels at controlled TWER. We find a strong correlation between TWER and WER of downstream ASR models.

10.21437/Interspeech.2023-1330

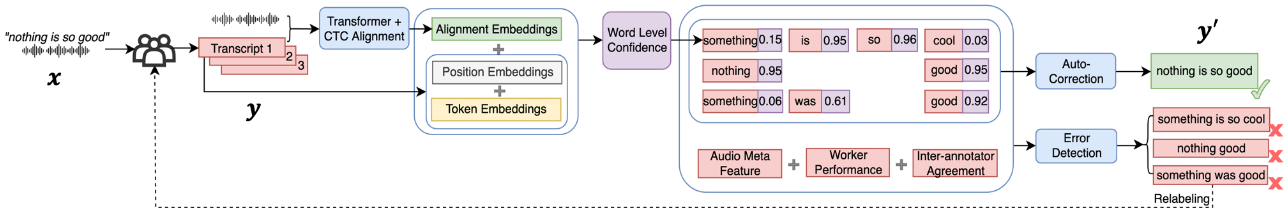

Figure 1: *The workflow of proposed ML-in-the-loop human transcription pipeline, where multiple crowdsourced workers transcribe one audio at a time. The audio-text alignment module generates the alignment embeddings, which are the input of a transformer to predict word-level confidence scores. The aligned transcripts among multiple transcripts and those confidence scores are used for auto-correction. Along with other meta features, they are fed to a model to predict sentence level likelihood of transcription error.*

The quality-improved training data can provide over 10% WER reduction. In summary, the contributions of this paper are:

1. We propose a new ML-in-the-loop data collection mechanism to produce high-quality transcriptions for ASR model training. Experiment shows it can significantly improve the human transcription quality.
2. We analyze the impact of transcription error on training ASR models, with incorrectly transcribed data by real human. We find a strong correlation between TWER and the WER of downstream ASR models.
3. We collect and release LibriCrowd[1], a large-scale public dataset of crowdsourced audio transcriptions. It contains 100 hours of English speech transcribed by 4433 human transcribers. The labeling cost is only $6 per speech hour. We believe LibriCrowd can benefit the research community to develop new MSR error correction models and design robust ASR models trained on noisy data.

2. Data Collection

This section introduces the new data collection mechanism we proposed. We collect speech transcription data via crowdsourcing from Amazon MTurk. Compared with the existing industry-level data pipelines [16, 17, 21], our data collection approach does not require expensive worker training, machine pre-labeling, human auditing, and annotator behavior monitoring. The transcription quality is controlled by the proposed confidence estimation modules (CEMs) at the labeling stage and error correction models (ECMs) at the post-labeling stage.

2.1. Data Source

In this work, we start from raw audios and collect human transcriptions to build the speech corpus. Our setting differs from many existing speech corpora [3, 15, 22] that were built on text-to-speech reading, since prepared scripts are rare in real-world speech applications. The audio files are collected from the LibriVox project, in which ground truth reference is available for us to measure the crowdsourced transcription quality.

2.2. Transcription Task Design

User Interface and Instruction: The UI is designed based on MTurk's audio transcription template[2]. Each task contains speech recordings loaded from external sources and blank fields for human annotators to provide transcription inputs. We increase the space between the text field and submit button to avoid incomplete submissions. To align with the format of ground truth transcription, we set instructions such as ignoring punctuation marks, transcribing digits as words, etc. Annotators are allowed to put a question mark at the positions where they are uncertain, instead of random guess.

Reward per Utterance: The trade-off between labeling cost, quality and efficiency has been studied in [23, 24, 25]. In many real-world applications the cost per utterance ranges from 0.5 to 5 cents, and it will mainly affect the efficiency rather than quality. Hence in this study we set reward at $0.01 per task, each of which contains five utterances.

Worker Selection: We do not set entrance exam and training requirement, unlike many existing industry-level transcription pipelines. All workers with the approval rate $\geq 95\%$ are qualified. The purpose is to reduce cost and increase diversity. Later the confidence estimation models we build will detect those low-quality submissions and send to a different worker for relabeling. We use the model results to filter out malicious or careless workers.

Number of Workers per Utterance: Each audio recording is assigned to five different workers. The inter annotator agreement will be an important signal for our confidence estimation afterwards. Our error correction models aggregate all transcripts to generate the final transcription.

3. Quality Improvement

We develop confidence estimation model (CEM) and error correction model (ECM) for quality improvement as soon as we collect every human transcription, as shown in Figure 1. Transcribers listen to the audio $\boldsymbol{x}$ and produce raw transcriptions $\boldsymbol{y}$. CEM detects error for every word in $\boldsymbol{y}$ and sends low-quality utterances for relabeling. ECM automatically corrects error words in $\boldsymbol{y}$ and generates transcriptions $\boldsymbol{y'}$ as the final output.

3.1. Confidence Estimation

Confidence scores are calculated at word and utterance level to assess the MSR quality. Various factors are considered, such as spell and grammar error, audio-text mismatch, transcriber's recent performance record, speech length and complexity.

CEM takes paired audio-text sequence $(\boldsymbol{x}, \boldsymbol{y})$ as input, and generates a confidence score c_i for each word $y_i \in \boldsymbol{y}$. We extract speech representations by pre-trained Wav2Vec2 [8] on LV-60k, and then use CTC-Segmentation [26] to align each word with its corresponding audio segment. Suppose $\boldsymbol{x}_{u_i:v_i}$ is the audio segment of the i^{th} word y_i, the alignment score s_i is calculated based on the emission probability $P(y_i|\boldsymbol{x}_{u_i:v_i})$ and CTC loss [5] given its bidirectional context.

$$\begin{aligned} \log s_i(\boldsymbol{x}, \boldsymbol{y}) &= \log P(y_i|\boldsymbol{x}, \boldsymbol{y}_{-i}) = \log P(y_i|\boldsymbol{x}_{u_i:v_i}) \\ &- \mathcal{L}_{CTC}(\boldsymbol{x}_{1:u_{i-1}}, \boldsymbol{y}_{1:i-1}) - \mathcal{L}_{CTC}(\boldsymbol{x}_{v_{i+1}:S}, \boldsymbol{y}_{i+1:T}) \end{aligned} \quad (1)$$

[1] https://github.com/GenerateAI/LibriCrowd
[2] https://github.com/GenerateAI/TransAudioUI

where $\boldsymbol{x}_{u_i:v_i} = (x_{u_i}, \cdots, x_{v_i})$, $\boldsymbol{y}_{1:i} = (y_1, \cdots, y_i)$, $\boldsymbol{y}_{-i} = (y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_T)$. $S = |\boldsymbol{x}|$ is the audio length, and $T = |\boldsymbol{y}|$ is the transcription length.

To further detect those errors from either syntactic or semantic perspective, we modify a BERT-like error detection model ELECTRA [27] by adding the alignment score s_i as the embedding. The discriminator of ELECTRA is fine-tuned on noisy transcriptions with real human errors. It is aligned with the ground truth reference and each word is labeled as correct or incorrect. The model outputs a confidence score c_i for each word y_i.

Then we train a model for utterance-level error detection. It predicts the expected number of word errors in an utterance based on the word-level CEM as well as meta features from three sources: (1) utterance complexity features such as audio duration, transcription length, and signal-to-noise ratio (SNR) estimated by the WADA-SNR algorithm [28]; (2) worker performance features such as task spending time, recent task accept rate; (3) inter annotator agreement, i.e., the edit distance from one transcription to other transcriptions for the same audio recording. The model is built upon gradient boosting trees and implemented by LightGBM [29]. The model provides the real-time feedback to transcribers by rejecting those unqualified responses and republishing the tasks.

3.2. Error Correction

Language models like BERT [30] and BART [31] have been widely used for text error correction. It is also well-studied to refine ASR outputs by second-pass rescoring on n-best hypotheses [32]. For MSR error correction, ECM can do word-level aggregation instead of sentence-level reranking in ASR error correction. This is because MSR outputs have independent sources while the n-best ASR hypotheses are generated by beam search decoding from the same acoustic model. However, unlike ASR system, human transcription does not have a confidence score generated by the first-pass decoder.

Therefore, we use the confidence score estimated by the word-level CEM. It measures the audio-text alignment as well as semantic consistency. Meanwhile, we perform text-text alignment between multiple human transcriptions of the same audio recording. The alignment is conducted by minimizing the edit distance between two text sequences using the Needleman–Wunsch Algorithm [33]. The correction is based on a simple voting scheme at each aligned position.

For a given utterance, suppose $w = y_i$ is the i^{th} word, $N(w, i)$ is the occurrence of word w at position i, N is the number of candidate transcriptions, and $c_i(w)$ is the confidence score. Then the overall scoring of word w is

$$\begin{aligned} s_i(w) &= \alpha N(w,i)/N + (1-\alpha) c_i(w) \\ w^* &= \arg\max_{w} s_i(w) \end{aligned} \tag{2}$$

where w^* is the selected word at position i. The trade-off parameter α between word frequency and confidence score is tuned on the training set. Note that when $\alpha = 1$, our algorithm degrades to ROVER [34]. Figure 1 includes an example for ECM.

4. Experiment

In this section, we evaluate the quality of MSR and analyze the impact of transcription error on ASR model training. MSR output and ASR output are compared with the ground truth reference to calculate TWER and WER, respectively.

Table 1: *LibriCrowd dataset statistics and raw TWER without quality improvement*

Subset	**# Utterances**	**speech hours**	**# Workers**	**# Responses**	**TWER (%)**
train-other-10h	3165	10.0	1258	18673	15.50
train-other-60h	17816	60.0	1136	20187	7.52
train-mixed-10h	2763	9.8	616	14231	5.97
dev-clean	2703	5.4	523	13994	6.10
test-clean	2620	5.4	527	13587	8.23
dev-other	2864	5.3	620	15235	12.69
test-other	2939	5.1	989	15950	16.61
all	34870	101.0	4433	111857	10.91

Table 2: *Word-level confidence estimation on LibriCrowd*

	Precision	Recall	F1
ELECTRA	0.6668	0.8316	0.7401
ELECTRA + Alignment	0.7590	0.8615	0.8070

4.1. Experiment Setup

We experiment on speech recordings with ground truth scripts to measure the quality of human transcription. Note that the ground truth is not available in most real scenarios. A surrogate is the consensus annotation of multiple expert transcribers.

For data source, we use audio recordings from LibriVox to create LibriCrowd. LibriVox[3] is a free public project that contains thousands of audio books read by a group of worldwide volunteers. The ground truth is directly from the text in the audio book. LibriCrowd contains 100 hours of English speech in three categories: (1) 70 hours audio recordings from higher-WER speakers (*train-other-10h*, *train-other-60h*), (2) 10 hours Limited Resource Training Set of Libri-Light [35] (*train-mixed-10h*), (3) Dev and Test Sets of LibriSpeech (20 hours). We release LibriCrowd to the research community under CC-BY-4.0 license. The statistics of LibriCrowd is summarized in Table 1.

The speech recordings in LibriCrowd are transcribed in a cold-start setting. For the first 10 hours of speech (*train-other-10h*), human adjudicators are hired to verify the submitted transcriptions. They listen to the audio and republish their transcriptions to replace the low-quality ones. This data is used to train CEM and ECM.

4.2. MSR Transcription Quality Analysis

Table 2 lists the performance of word-level CEM. Compared to the ELECTRA baseline, having the alignment embedding can improve F1 score by 9% on LibriCrowd. The audio-text alignment serves as a surrogate "ASR confidence" which provides a reliable confidence estimation for human transcription.

Table 3 shows the MSR transcription quality before and after applying the proposed quality improvement approach. In summary, CEM-based relabeling can reduce 22% TWER (10.91 → 8.48), and ECM further reduces 42% TWER (8.48 → 4.94). Together, our approach reduce 55% TWER from raw transcription (10.91 → 4.94). Note that our method is much better and cheaper than CrowdSpeech [21] (4.94 v.s 7.84) in which each recording was transcribed by seven workers.

At the labeling stage, the average relabeling rate is 5.4%. In those rejected transcriptions, deletion error dominates the TWER, while in the accepted responses, substitution error is prevalent (Table 4). We find some rejected samples are empty or incomplete. Through the email communication, some transcribers mentioned they mis-clicked the submit button before finishing the task. Then we slightly increased the gap between the submit button and the transcription text box. The reduced deletion error can reflect the TWER gap between those rejected

[3] `https://librivox.org/`

Table 3: *Comparison of proposed quality improvement methods with baselines. Quality is evaluated by TWER (%) against ground truth. A lower value means higher quality.*

	train-mixed-10h	dev-clean	test-clean	dev-other	test-other	average
Raw Transcription	5.97	6.10	8.23	12.69	16.61	10.91
CEM + Random	5.00	4.94	5.77	10.20	13.01	8.48
CEM + Longest	5.12	5.83	6.30	10.20	12.42	8.69
CEM + Best Worker	4.60	4.72	5.17	9.51	11.30	7.68
CEM + RescoreBERT	4.11	4.58	5.96	9.99	11.14	7.16
CEM + ECM	**2.99**	**2.76**	**3.05**	**6.45**	**7.51**	**4.94**
CEM + Oracle	2.05	1.69	1.88	4.13	5.00	3.18
CrowdSpeech + ROVER	N/A	6.76	7.29	13.19	13.41	10.16
CrowdSpeech + T5	N/A	N/A	5.22	N/A	10.46	7.84
CrowdSpeech + Oracle	N/A	3.81	4.32	8.26	8.50	6.22

Table 4: *Error types in raw MSR transcription*

Accepted	train-mixed-10h	dev-clean	test-clean	dev-other	test-other	average
Length (# word)	35	20	20	18	18	22
Deletion (%)	0.87	0.91	1.17	1.67	2.77	1.48
Insertion (%)	0.44	0.42	0.48	0.95	1.33	0.72
Substitution (%)	3.58	3.68	4.12	6.90	8.34	5.32
TWER (%)	4.89	5.01	5.78	9.52	12.43	7.53
Rejected						
Length (# word)	17	12	9	9	9	11
Deletion (%)	44.72	33.46	57.22	48.94	49.31	46.73
Insertion (%)	1.74	2.71	2.43	3.54	3.91	2.87
Substitution (%)	15.40	17.49	15.12	20.47	19.34	17.56
TWER (%)	61.85	53.66	74.77	72.95	72.56	67.16

transcriptions in *test-clean* and *dev-clean*. This finding shows a good UI design can improve human transcription quality.

When we published the tasks we followed in the order of *train-other-10h*, *test-other*, *dev-other*, *test-clean*, *dev-clean*, and *train-mixed-10h*. Through that process we observed the trend of improving worker performance and decreasing worker number. It shows CEM-based quality assessment can effectively eliminate invalid task submissions and help filter out poorly-performed workers. Meanwhile, the frequency of substitution error reduced (17.56 → 5.32), thanks to the alignment embedding that detects audio-text mismatch.

At the post-labeling stage, we compare the proposed ECM with five baselines:

- Random: randomly pick a transcription as the final output
- Longest: pick the transcription with the longest text length
- Best worker: pick the transcription from the best worker
- Oracle: pick the transcription with the lowest edit distance from the ground truth reference
- RescoreBERT [32]: a BERT-based second-pass rescoring model for ASR error correction. We directly use it to rerank MSR candidates by equally setting the first pass scores to 0.5.

ECM is trained on *train-other-10h*. We consider both word frequency and its confidence score, which detects acoustic and semantic defects based on the alignment of audio and text. The optimal trade-off parameter is learnt at 0.8. Table 3 shows our model outperforms the baselines and RescoreBERT by over 30%. Note that the voting in our method is at word level while RescoreBERT is reranking the entire sentence, so ideally our method may exceed the performance of Oracle.

4.3. Transcription Error Impact on ASR Model Training

This section investigates the impact of transcription error on ASR model training. We fine-tune two pre-trained ASR models Wav2Vec2 and WavLM on noisy data, and evaluate their performance on test sets. For fine-tuning, we use the same setup as in [8, 11]. The ASR models are decoded by a 4-gram LM with decoding parameters tuned on the noisy dev set. Table 5 shows the result of Wav2Vec2 trained on *train-mixed-10h*. Our quality improvement method provides a relative WER reduction of 9.52% for the ASR system. (R_WER: 14.49% → 4.97%)

Table 5: *Impact of training data quality on ASR performance*

	label	TWER	WER (w/o LM)	R_WER (w/o LM)	WER (w/ LM)	R_WER (w/ LM)
test-clean	Raw Transcription	5.97	10.71	12.97%	5.53	14.49%
	CEM + Random	5.00	10.27	8.33%	5.35	10.77%
	CEM + ECM	**2.99**	**9.69**	**2.22%**	**5.07**	**4.97%**
	Ground Truth	0.00	9.48	0.00%	4.83	0.00%
test-other	Raw Transcription	5.97	19.03	8.37%	12.01	12.56%
	CEM + Random	5.00	18.18	3.53%	11.64	9.09%
	CEM + ECM	**2.99**	**17.65**	**0.51%**	**11.01**	**3.19%**
	Ground Truth	0.00	17.56	0.00%	10.67	0.00%

To quantify the relation between training data quality and ASR model's performance, we set TWER at predefined scales (0 - 10%) by randomly mixing noisy human transcriptions with ground truth reference. We do not use any data augmentation strategy like SpecAugment [36]. Figure 2 shows a strong correlation between TWER and R_WER.

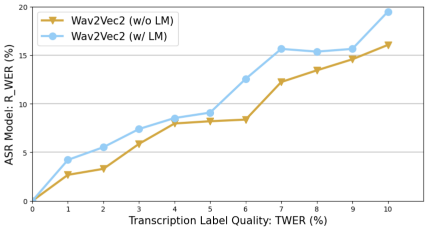

Figure 2: *ASR models fine-tuned at controlled noise level*

We further investigate whether the impact of transcription error can be mitigated by switching to a larger model with more training data. We fine-tune WavLM on clean and noisy data (TWER = 5%) with training data size increased from 1 to 80 hours. The asymptote gap in Figure 3 indicates increasing data and model size cannot mitigate the harm of transcription defect.

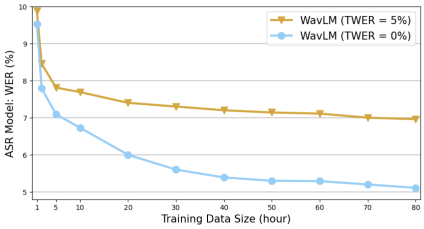

Figure 3: *ASR models fine-tuned at controlled training data size*

5. Conclusion and Future Work

In this paper, we propose a new speech transcription data collection method with ML-in-the-loop quality improvement mechanisms to detect and correct human errors. In particular, CEM detects low-quality transcriptions for relabeling, and ECM automatically corrects word errors and generates the final transcription. We investigate the impact of transcription error on ASR model training, and find a strong correlation between label quality and ASR system's performance.

We collect and release LibriCrowd, a large-scale crowdsourced dataset of human speech transcription. We apply the proposed quality improvement method on the collected human transcriptions. Our method reduces TWER by over 50%, which gives 10% relative WER reduction to the ASR models trained on it. We believe this dataset can benefit the research community to develop advanced MSR error correction models, as well as to design ASR algorithms that are robust on noisy data.

For future work, we will continue to collect crowdsourced data from scripted and unscripted speech to enhance LibriCrowd. We will explore more transcription error mitigation methods such as including large language models.

6. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, 2017.

[2] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach.* Kluwer Academic Publishers, 1993.

[3] J. S. Garofolo and et al., "Timit acoustic-phonetic continuous speech corpus," *Philadelphia: Linguistic Data Consortium*, vol. LDC93S1, 1993.

[4] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. Interspeech 2017*, 2017, pp. 498–502.

[5] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural nets," in *ICML*, 2006.

[6] A. Graves, "Sequence transduction with recurrent neural networks," *CoRR*, vol. abs/1211.3711, 2012.

[7] T. Brown and et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[9] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *ASRU*, 2021, pp. 244–250.

[10] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[11] S. Chen and et al., "Wavlm: Large-scale self-supervised pretraining for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[12] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP*, 2020, pp. 7829–7833.

[13] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[14] W. Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu, "ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context," in *Proc. Interspeech 2020*, 2020, pp. 3610–3614.

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[16] M. Liu, C. Zhang, H. Xing, C. Feng, M. Chen, J. Bishop, and G. Ngapo, "Scalable data annotation pipeline for high-quality large speech datasets development," *CoRR*, vol. abs/2109.01164, 2021. [Online]. Available: https://arxiv.org/abs/2109.01164

[17] M. Levit, Y. Huang, S. Chang, and Y. Gong, "Don't count on asr to transcribe for you: Breaking bias with two crowds," *Interspeech*, August 2017.

[18] R. Sundaram and J. Picone, "Effects on transcription errors on supervised learning in speech recognition," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2004, pp. I–169.

[19] I. Chen, B. King, and J. Droppo, "Investigation of training label error impact on RNN-T," *CoRR*, vol. abs/2112.00350, 2021. [Online]. Available: https://arxiv.org/abs/2112.00350

[20] M. K. Odell, "The profit in records management," *Systems (New York)*, vol. 20, p. 20, 1956.

[21] N. Pavlichenko, I. Stelmakh, and D. Ustalov, "Crowdspeech and vox DIY: benchmark dataset for crowdsourced audio transcription," in *NeurIPS Datasets and Benchmarks*, 2021.

[22] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-Âmultilingual speech corpus," in *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222.

[23] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 5270–5273.

[24] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *NAACL*. Los Angeles, California: Association for Computational Linguistics, Jun. 2010, pp. 207–215.

[25] K. Audhkhasi, P. Georgiou, and S. S. Narayanan, "Accurate transcription of broadcast news speech using multiple noisy transcribers and unsupervised reliability metrics," in *ICASSP*, 2011.

[26] L. Kürzinger, D. Winkelbauer, L. Li, T. Watzel, and G. Rigoll, "Ctc-segmentation of large corpora for german end-to-end speech recognition," in *Speech and Computer*. Springer International Publishing, 2020, pp. 267–278.

[27] H. Futami, H. Inaguma, M. Mimura, S. Sakai, and T. Kawahara, "Asr rescoring and confidence estimation with electra," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 380–387.

[28] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proc. Interspeech 2008*, 2008, pp. 2598–2601.

[29] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Neural Information Processing Systems*, 2017.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," in *NAACL*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[31] M. Lewis and et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.

[32] L. Xu, Y. Gu, J. Kolehmainen, H. Khan, A. Gandhe, A. Rastrow, A. Stolcke, and I. Bulyko, "Rescorebert: Discriminative speech recognition rescoring with bert," in *ICASSP*, 2022, pp. 6117–6121.

[33] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[34] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 1997, pp. 347–354.

[35] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP*, 2020, pp. 7669–7673.

[36] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.