Interspeech 2022
18-22 September 2022, Incheon, Korea

Reducing Domain mismatch in Self-supervised speech pre-training

Murali Karthick Baskar [Φ], *Andrew Rosenberg* [†], *Bhuvana Ramabhadran* [†], *Yu Zhang* [†]

[Φ] Brno University of Technology, [†] Google Inc.,
baskar@fit.vutbr.cz, {rosenberg,bhuv,ngyuzh}@google.com

Abstract

Masked speech modeling (MSM) methods such as wav2vec2 or w2v-BERT learn representations over speech frames which are randomly masked within an utterance. While these methods improve performance of Automatic Speech Recognition (ASR) systems, they have one major limitation. They treat all unsupervised speech samples with equal weight, which hinders learning as not all samples have relevant information to learn meaningful representations. In this work, we address this limitation. We propose ask2mask (ATM), a novel approach to focus on specific samples during MSM pre-training. ATM employs an external ASR model or *scorer* to weight unsupervised input samples by performing a fine-grained data selection. ATM performs masking over the highly confident input frames as chosen by the scorer. This allows the model to learn meaningful representations. We conduct fine-tuning experiments on two well-benchmarked corpora: LibriSpeech (matching the pre-training data) and, AMI and CHiME-6 (not matching the pre-training data). The results substantiate the efficacy of ATM on significantly improving the recognition performance under mismatched conditions while still yielding modest improvements under matched conditions.
Index Terms:Self-supervision, Wav2vec2, pretraining, Data selection, Domain mismatch, asr, speech recognition

1. Introduction

Self-training and self-supervised training techniques rely on huge amounts of unlabeled speech or text data for better generalization. The self-training techniques such as pseudo-labeling [1, 2] and student-teacher training [3] have shown promising improvements by incorporating the data selection process. This data selection step removes pseudo-labels with less confidence as denoted by the teacher model before feeding the input to a student model. Xu et al. [4] show that self-training and self-supervised training are complementary to each other and also show that self-supervised models act as good initialization for self-training techniques.

Masked speech modeling (MSM) is the recent and successful self-supervised learning technique, thanks to the advent of BERT [5] in NLP which inspired learning speech representations from masked inputs. MSM techniques such as wav2vec2 [6], HuBERT [7] and w2v-BERT [8] have shown considerable gains across various down-stream speech tasks and have become the go-to modeling approaches for ASR.

However, MSM does not have a data selection scheme to discard the irrelevant input samples and instead imposes burden on the training criterion to learn the relevance of the input samples in learning meaningful representations. [9] noticed the impact of not selecting relevant data from the huge amounts of unsupervised data during pre-training by showing degradation in ASR performance when fine-tuned to a target dataset with limited data. To mitigate this constraint, [10] introduced substantially more fine-tuning data related to the target dataset but did not achieve satisfactory results. [9] attempted to address this issue by heuristically selecting the data from a closed set of unsupervised speech databases or by including data relevant to target dataset along with the existing pre-training dataset. However, this data selection approach is not done within the existing pre-training dataset and it is not completely empirically motivated.

In this study, we propose a simple strategy named *ask2mask (ATM)* to incorporate data selection within a chosen pretraining dataset. Here, the masking is done over the speech frames with higher confidence as determined by the *scorer*. This is contrary to the random selection of frames to be masked in conventional MSM models. We hypothesize that this guided selection of frames to be masked allows the model to focus on the frames which can provide meaningful representations. The scoring model used in this work is a speech recognition model trained on small amount of data and provides frame-level confidence for each input.

In [11], phonetic knowledge is injected to mask over phonetic segments to perform spectral augmentation. Deterministic masking schemes are also proposed interms of phonemes in [12] and using Entropy in [13] to improve multiple downstream speech tasks. Our ATM approach is primarily motivated based on the recent work by [14] on semi-supervised learning of conventional ASR systems which shows that performing data selection at frame-level or token-level on unsupervised data provides better performance. Few works on unsupervised learning also highlight the importance of weighting the data based on its confidence [15, 16, 17]. We hypothesize that ATM can leverage the effect of data selection within a particular training corpus to further enhance the recognition performance of MSM techniques.

To summarize, our contributions are listed as follows:

- *Novelty*: To the extent of our knowledge, ATM is the first approach to incorporate a within-corpus data selection strategy in MSM. We also show that data selection can be simply performed inside MSM by guided selection of frames to be masked using a scorer model.
- *Technical contributions*: We provide a simple strategy to incorporate data selection into MSM pretraining by applying the confidence of the scorer. ATM is designed to be compatible to all MSM based pre-training techniques.
- *Empirical study*: Analysis is done to find an optimal masking percentage for ATM and we highlight the effectiveness of ATM across varying masking percentages. The importance of masking frames with high confidence is substantiated by empirically comparing it with masking low confident frames and random frames respectively. Experiments are performed on AMI and CHiME-6 data which is from a distinct condition compared to Libri-light corpus used for MSM based pretraining.

10.21437/Interspeech.2022-736

2. Ask2Mask (ATM)

The primary reason to employ pre-training models is to exploit the abundantly available unsupervised data for improving ASR under limited availability of supervised data. The MSM models such as wav2vec2 [6] and w2v-BERT [18] learn from unsupervised data but they treat each data sample with equal weight for computing the final objective. For instance, let the input speech sequence $\mathbf{X} = [x_1, x_2, ..., x_{T'}]$, where x_t is the log Mel-filterbank feature vector at time t. $\mathbf{X}$ is sent to the feature encoder Φ to obtain the encoded representations $\mathbf{E} = \Phi(\mathbf{X})$. The feature encoder contains convolutional layers performing subsampling at a factor of 4 and reducing the total number of frames of an utterance from $T^{'}$ to T to get $\mathbf{E} = [e_1, e_2, ..., e_T]$. $\mathbf{E}$ is then sent to two parallel modules: 1) masking component, and 2) quantizer. The masking is done over sets of frames or blocks $b_1, b_2, ..., b_K$ and accommodates overlap between blocks. Here K is the number of masked blocks in a randomly masked encoded sequence $\tilde{\mathbf{E}}$. The block $b_k = [i_k, c]$, where i_k is the starting index of the masked block and c is the corresponding right context size denoting the number of consecutive speech frames. Here i_k are randomly sampled from a uniform distribution. It has been empirically observed by [6] that 49% of the frames are masked and $c = 10$ is chosen as the best hyperparameter value to attain best representation during pre-training.

Instead, we generate a score s_t for each encoded frame e_t. This is used to select relevant data in a fine-grained manner during masking for computing the loss objective. Here, we hypothesize that pre-training with data that closely resembles the target domain leads to better recognition performance after fine-tuning. Finally, a Gumbel-softmax quantizer component is used to get quantized representations which act as targets for wav2vec2 and w2v-BERT models.

2.1. Methodology

For each encoded feature frame $e_t \in \mathbf{E}$, the scorer emits probabilities $p(v_t = l \,|\, \mathbf{E}); l \in \mathbf{L}$ of the frame belonging to a particular label. The scorer model is a CTC based frame-synchronous ASR model separately trained with a limited amount of data. Finally, the confidence score s_t of the frame is defined as the maximum probability across all labels:

$$s_t = \max_l p(v_t = l \,|\, \mathbf{E}) \tag{1}$$

We sample K masking start indices $\{i_1, .., i_k\}$ with probabilities without any replacement. That is, we sample beginning frames with probability proportional to the scores of each frame. This is the key difference between ATM and other masking approaches. Prior works uniformly sample the start indices of each masking block $b_{1:K}$, while the ATM uses the probability distribution induced by the scorer. K is determined by the percentage of frames to be masked.

We hypothesize that frames with maximum confidence from an external scoring model will be 1) easiest to learn using an MSM training criteria and 2) most informative in for pretraining to facilitate fine-tuning. Conversely, the lowest confidence frames, those more confusable to an external scoring model, will be the least reliably learned by MSM and least informative for pretraining.

The resulting frames are sent as input to the MSM architecture and the final loss objective $\mathcal{L}$ is determined by either of the MSM objectives such as wav2vec2: $\mathcal{L}_{wv}$ or w2v-bert: $\mathcal{L}_{wb}$.This modified training objective allows the model to focus on learning from gradients calculated from the frames with high confidences.

3. Experimental Setup

All experiments including pre-training and fine-tuning are performed using 80 dimensional log Mel-filterbank features computed over the sampled 16kHz audio. Datasets (such as AMI) contains wideband audio and are downsampled to 16kHz. We evaluate with the test-other (LibriSpeech partition) to show the importance of ATM on matched data conditions, while IHM-eval and SDM-eval (AMI partitions) is used to validate the model under mismatched conditions.

3.1. Datasets used

Pretraining (PT): Libri-light (LL-60k) dataset contains 60k hours of unlabeled speech and is used to pre-train all MSM models. LL-60k is the most widely used large unsupervised speech corpus for various PT techniques. Each input speech sequence is constructed by first randomly selecting 32-64 seconds segments from the original utterance. From these segments, a contiguous 32 second region is extracted from a random starting point on-the-fly during MSM PT as described in [19].

Finetuning (FT): Different target datasets including 1) 100 hrs of Librispeech (LS-100) [20]. 2) 100 hours of AMI and 3) speechstew (approx. 5k hours) [10] are used to perform our FT experiments. Each dataset used is specific to a certain target data condition, for instance LS-960 is closely matches the LL-60k, AMI dataset is distinct from the LL-60k condition and it contains speech from two kinds of microphones (i) Independent head microphone (IHM). (ii) single distant microphone (SDM). SpeechStew is composed of datasets chosen from multiple conditions to create a mixed domain aggregate corpus. Processing details are described in [10].

Evaluation: We evaluate ATM performance on AMI using IHM-eval and SDM-eval. Finally, we also evaluate using CHiME-6 [21] without using any FT data from CHiME-6 training set to compare the performance of ATM on completely unseen target dataset.

Scorer training data: A CTC [22] based conformer model with 100M parameters is trained on the fine-tuning data, thereby not requiring additional supervised data. The scorer is chosen based on the target downstream task and in addition to this, the scorer needs to be frame-synchronous to provide confidence for every frame in a speech sequence. In this work, we use a frame-synchronous ASR system as the scorer by employing the connectionist temporal classification (CTC) objective. The CTC is preferred over the RNN-T by analysing the reliability of the frame-level predictions. Word-piece model (WPM) with 1024 tokens are used as labels for training the scorer models.

3.2. MSM architecture

W2v2-cfr: This is a wav2vec2 with conformer based context network which first encodes the filterbank features using two 2D convolutional layers with strides (2,2). Model has 100M/600M parameters and is denoted as "w2v2-cfr-L/XL". HuBERT-cfr-L/XL is similar to w2v2-cfr-L/XL - it differs in using the k-means based quantizer with 1024 targets and computes the cross-entropy loss as described in [7]. The "L/XL" size models contains context network Ω 12/24 conformer layers with 8 attention heads and 1024 hidden dimensions.

W2v-BERT: W2v-BERT is explored using two model sizes: one with 100M parameters denoted as "w2v-BERT-L" and containing 2 conformer layers in context net Ω and 4 conformer layers in Λ. A 600M parameter model is denoted as "w2v-BERT-XL" contains 8 conformer layers in Ω and 24 conformer

layers in Λ. Each conformer block contains 1024 hidden dimensions with 8 attention heads, kernel size of 5 with local context of 128. The remaining architecture is identical to the configuration defined in [8].

3.3. PT and FT configuration

The models L/XL are trained with a global batch size of 512/2048 on 64/256 Google TPU V3 cores for 2-4 days respectively. Adam optimizer is used with a learning rate schedule (Section 5.3 of [23]) with 2e-3 as peak learning rate and 25k warmup steps. The model training configuration follows similar procedure as described in [19].

The FT is done by employing the context network from the PT model by adding a final layer with 1024 WPM units learnt using the RNN-T objective function [24]. The FT is done on w2v-BERT-XL, w2v2-cfr-XL and HuBERT-cfr-XL after 400k PT model updates. The w2v-BERT-L model is FT after 900k PT model updates. w2v-BERT-L is used to initially perform wide range of analysis and hyper-parameter optimization on ATM. w2v-BERT-XL is finally used to compare the results of ATM across existing works in literature. w2v2-cfr-XL and HuBERT-cfr-XL are also used in our experiments. All these models are trained with the same configuration as in [19].

4. ATM analysis

The empirical study on ATM is done primarily using w2v-BERT-L since this generates the best WER performance across similarly sized models (cf. Figure 2). The pre-trained models are fine-tuned with either LS-100 or AMI. The resulting finetuned models are evaluated on IHM and SDM evaluation sets to understand the domain generalization aspect of ATM. Librispeech evaluation sets are used in unison to study how ATM behaves under matching domain condition. We initially conduct our analysis on choosing the optimal scorer and to understand how much supervision is required. Three different scorers are trained using 1 hour, 10 hours and 100 hours of Librispeech data. The scorers are evaluated using 1 hour of Libri-light data and the results are in table 1. We chose 1 hour of Libri-light data since the pretraining data is from the same domain (Libri-light).

Table 1: *Comparison between scorers trained with different amount of supervised data (1h, 10, 100h of Librispeech). The scorer is evaluated on 1hour of Libri-light data.*

#Hours.	% WER
1	45.4
10	27.6
100	25.8

4.1. Masking percentages

The number of masked frames within an utterance plays a key role in masked input learning and in this study, we vary the masking percentages from 30% to 50% to determine the best percentage for ATM approach. Previous works on wav2vec2 [6] showed that masking 49% of the frames is ideal for 30 second utterance and this has been followed subsequent works such as HuBERT and w2v-BERT. In case of ATM, this can differ as the frames selected are of higher confidences. Figure 1 shows that ATM achieves its "sweet spot" with 40% masking for both IHM-eval and test-other set. Interestingly ATM's performance

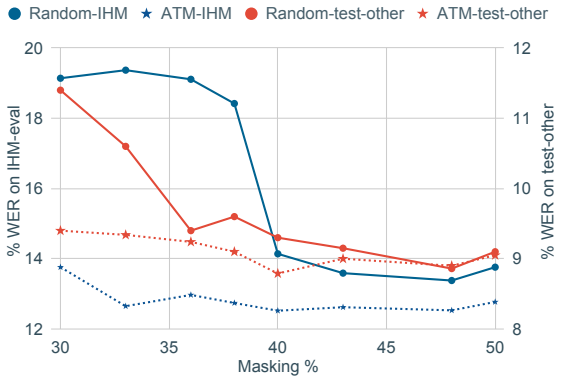

Figure 1: *Recognition performance of w2v-BERT with ATM and random masking on IHM-eval and test-other sets by varying the masking percentage during pre-training. The FT is performed on LS-100 for evaluating test-other, while IHM-eval is evaluated with model FT with AMI. Random masking shows a substantial shift in performance when varying the masking from 30% to 40%, while the ATM remains robust to changes in masking percentage.*

is stable across large variations in masking rates with relatively good performance with masking rate as low as 30%. This is a significant difference from the uniform sampling of prior work which suffers significant drop in performance as the masking rate goes below 40%. The result indicates that masking the right set of frames, which ATM aims to do, is able to promote more stable performance. For instance, ATM achieves a %WER of 12.65 with 33% masking and 12.52 with 40% masking on IHM-eval respectively as shown in Figure 1. The recognition performance on test-other and IHM-eval improves over baseline from 8.86% to 8.79% and 13.38% to 12.52% respectively by using ATM.

4.2. Consistency across different architectures

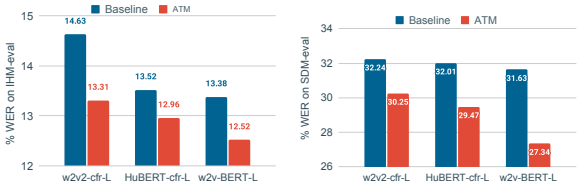

Figure 2: *Performance comparison of different MSM architectures with and without applying ATM on IHM-eval and SDM-eval in AMI. All these models are FT using AMI. Here "cfr" refers to conformer.*

Figure 2 shows that ATM consistently outperforms on both IHM-eval and SDM-eval across multiple MSM architectures including wav2vec2 and HuBERT. In the case of IHM-eval, ATM attains a relative improvement of 9% over w2v2-cfr-L, 4% relative improvement over HuBERT-cfr-L and 5% relative gain over w2v-BERT-L baseline models respectively. W2v2-cfr-L using ATM obtained 6.2% relative improvement over its baseline counterpart and HuBERT-cfr-L with ATM attained 7.9% rel. improvement over HuBERT-L baseline on SDM-eval respectively. On the other hand w2v-BERT-L baseline is better compared to w2v2-cfr-L and HuBERT-cfr-L on both IHM-eval and SDM-eval by achieving 12.52% and 27.34% WER respectively.

4.3. Analysis on Librispeech

Experimental analysis is conducted using different model architectures to validate the effect ATM on LS-100 and are present in table 2. The impact of increasing the model parameters from "L" size to "XL" size, we FT on LS-100 using MSM models with XL size and the results are in table 2. We did not find any consistency in the performance across the evaluation sets using any of the MSM architectures. Slight gains are observed on test or dev or dev-other using w2v2-cfr-XL. Once the baseline in w2v-BERT-XL gets better, ATM did not achieve gains on test-other. This scenario can be explained due to effectiveness of MSM pre-training under matched condition and can perform well without any necessary data selection approach.

Table 2: *Performance comparison of different MSM architectures with and without applying ATM on all evaluation sets on Librispeech.*

Model	PT-LL, FT-LS100			
	dev	dev-other	test	test-other
w2v-BERT-L	3.78	8.86	3.85	9.32
+ATM	3.71	8.97	3.89	8.92
w2v2-cfr-XL	2.5	4.7	2.6	4.9
+ATM	2.4	4.6	2.5	5.0
HuBERT-cfr-XL	2.5	4.7	2.6	5.0
+ATM	2.5	4.6	2.5	5.0
w2v-BERT-XL	2.4	4.4	2.5	**4.6**
+ATM	**2.3**	**4.4**	**2.4**	4.7

5. Results

In this Section, the XL models are used to compare the importance of ATM on AMI and CHiME-6. These three datasets show the effect of ATM on diverse conditions with a much larger model. Results are compared with appropriate prior work. Table 3 presents the results of ATM on AMI by comparing it with w2v2-conformer-XL baseline and w2v-BERT-XL baselines. We include w2v2-conformer-XL to further test the consistency of ATM on XL models when evaluated on harder tasks. ATM observes consistent gains over baseline on both IHM-eval and SDM-eval when trained with XL models.

Table 3: *%WER obtained by FT with AMI using w2v-BERT-XL model using baseline and ATM. Evaluation is done on AMI test sets to highlight the effect on mismatched condition.*

MSM arch.	IHM-eval	SDM-eval
w2v2-cfr-XL	10.4	25.7
+ATM	10.0	24.5
w2v-BERT-XL	10.1	25.1
+ATM	9.5	23.7

Table 4 analyses the effect of ATM on multiple evaluation sets such as AMI and CHiME-6. These four sets are chosen based on the mismatch range from minimum to maximum and for instance, Commonvoice has the minimum mismatch with Libri-light data, while CHiME-6 has the maximum mismatch. The state-of-the-art results published in [10] are obtained by choosing the best Conformer model supervisedly trained with multiple datasets such as AMI, CommonVoice, Broadcast News, Librispeech, Switchboard/Fischer, TED-LIUM and Wall Street Journal. Note that the training data did not include the CHiME-6 data. The authors in [10] show that simply training an ASR with lots of data leads to best results compared to the wav2vec2 finetuned model. Their best results are denoted in table 4 and will be used to compare with our best ATM results.

Table 4: *Comparison with state-of-the-art results on SpeechStew. The FT is done on SpeechStew and the results are evaluated using Kaldi scoring to match published results. Note that the model has never seen any CHiME-6 data, and we use it as an example for **zero-shot** learning mode on how the model performs on chime-6 wihout seeing any of its training data.*

Model	AMI		CHiME-6
	IHM	SDM	
Speechstew [10]	9.0	21.7	57.2
w2v2-cfr-XL [10]	9.6	23.8	56.4
w2v-BERT-XL	9.2	21.5	55.5
+ ATM	9.0	21.0	54.3

Our baseline w2v-BERT-XL attained better results over the published w2v2-conformer-XL and Speechstew results. In Commonvoice and CHiME-6, the baseline attained 7.4% and 2.9% relative improvement over Speechstew respectively. However, by including our ATM with w2v-BERT-XL, there was consistent improvement across all range of mismatched domains. This result clearly justifies that selection of reasonable input samples during pre-training reduces the necessity of having finetuning data from the same domain to improve performance. To further substantiate this, the results on AMI show a 4.6% relative improvement on AMI-SDM over Speechstew which is of different domain compared to pre-training domain. In case of minimal mismatch domain such as Commonvoice, the ATM attained 11.6% relative improvement over Speechstew. These observations show that ATM demonstrate their effectiveness to generalize to unseen and challenging speech recognition conditions.

6. Conclusion

In this work, we introduce ask2mask (ATM) to perform data selection over unsupervised samples for MSM based pre-training to focus on relevant information and learn meaningful representations. ATM achieves 21.0% WER on mismatched AMI SDM set with guided masking. We empirically show that ATM is more robust to changes in masking percentage compared to random masking. as typically used in MSM. Our results substantiate the importance of learning from high confident frames by attaining improvements across multiple evaluation sets. An important aspect of ATM approach is its flexibility to incorporate into any MSM pretraining techniques. In our future work, we wish to apply ATM over pretraining data containing data from multiple domains [9, 25] to achieve further improvements. We also consider two future enhancements to ATM: (1) Joint training of the scorer model with MSM model by simultaneous training on supervised and unsupervised data. (2) Perform active learning by sharing the parameters of MSM with the scorer once the MSM is well trained.

7. References

[1] H. Scudder, "Probability of error of some adaptive pattern-recognition machines," *IEEE Transactions on Information Theory*, vol. 11, no. 3, pp. 363–371, 1965.

[2] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *ICASSP 2020)*, pp. 7084–7088, IEEE, 2020.

[3] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," *arXiv preprint arXiv:2005.09629*, 2020.

[4] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau, R. Collobert, G. Synnaeve, and M. Auli, "Self-training and pre-training are complementary for speech recognition," in *ICASSP 2021)*, pp. 3030–3034, IEEE, 2021.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[7] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "Hubert: How much can a bad teacher benefit asr pre-training?," in *ICASSP 2021*, pp. 6533–6537, IEEE, 2021.

[8] 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," *arXiv preprint arXiv:2108.06209*, 2021.

[9] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, *et al.*, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *arXiv preprint arXiv:2104.01027*, 2021.

[10] W. Chan, D. Park, C. Lee, Y. Zhang, Q. Le, and M. Norouzi, "Speechstew: Simply mix all available speech recognition data to train one large neural network," *arXiv preprint arXiv:2104.02133*, 2021.

[11] C. Wang, Y. Wu, Y. Du, J. Li, S. Liu, L. Lu, S. Ren, G. Ye, S. Zhao, and M. Zhou, "Semantic mask for transformer based end-to-end speech recognition," *arXiv preprint arXiv:1912.03010*, 2019.

[12] X. Yue and H. Li, "Phonetically Motivated Self-Supervised Speech Representation Learning," in *Proc. Interspeech 2021*, pp. 746–750, 2021.

[13] Y. Levine, B. Lenz, O. Lieber, O. Abend, K. Leyton-Brown, M. Tennenholtz, and Y. Shoham, "Pmi-masking: Principled masking of correlated spans," *arXiv preprint arXiv:2010.01825*, 2020.

[14] K. Veselý, L. Burget, and J. Cernocký, "Semi-supervised dnn training with word selection for asr.," in *Interspeech*, pp. 3687–3691, 2017.

[15] F. Wessel, R. Schluter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on speech and audio processing*, vol. 9, no. 3, pp. 288–298, 2001.

[16] Z. Ren, R. Yeh, and A. Schwing, "Not all unlabeled data are equal: Learning to weight data in semi-supervised learning," *NeurIPS*, vol. 33, 2020.

[17] C. Coleman, C. Yeh, S. Mussmann, B. Mirzasoleiman, P. Bailis, P. Liang, J. Leskovec, and M. Zaharia, "Selection via proxy: Efficient data selection for deep learning," *arXiv preprint arXiv:1906.11829*, 2019.

[18] 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," 2021.

[19] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2010.10504*, 2020.

[20] 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, IEEE, 2015.

[21] S. W. et. al., "Chime-6 challenge:tackling multispeaker speech recognition for unsegmented recordings," 2020.

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, pp. 369–376, 2006.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, pp. 5998–6008, 2017.

[24] 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, IEEE, 2020.

[25] T. Likhomanenko, Q. Xu, V. Pratap, P. Tomasello, J. Kahn, G. Avidov, R. Collobert, and G. Synnaeve, "Rethinking evaluation in asr: Are our models robust enough?," *arXiv preprint arXiv:2010.11745*, 2020.