
Proceedings Paper:
Park, C. orcid.org/0000-0001-6671-1671, Kang, H. and Hain, T. orcid.org/0000-0003-0939-3464 (2024) Character error rate estimation for automatic speech recognition of short utterances. In: Proceedings of 2024 32nd European Signal Processing Conference (EUSIPCO). 2024 32nd European Signal Processing Conference (EUSIPCO), 26-30 Aug 2024, Lyon, France. Institute of Electrical and Electronics Engineers (IEEE) , pp. 131-135. ISBN 9798331519773

https://doi.org/10.23919/eusipco63174.2024.10715433

Character Error Rate Estimation for Automatic Speech Recognition of Short Utterances

Chanho Park, Hyunsik Kang, Thomas Hain
Speech and Hearing Research Group, University of Sheffield
Sheffield, United Kingdom
{cpark12, hkang17, t.hain}@sheffield.ac.uk

***Abstract*—The quality of an automatic speech recognition (ASR) system's output can be measured by comparing it with a gold standard reference. Evaluating an error rate (ER) is costly and therefore not always possible. Instead, one can aim to provide estimates for quality, without explicit reference. Prior work has concentrated on confidence scoring or word error rate (WER) estimation. The latter is typically model based, and it was found that the performance of a WER estimation model degrades when it is trained on short utterances. To address this issue this work presents an ER estimation model using character error rate (CER), called Fe-CER. The ER estimation model for ASR system's output employs character-level tokenisation for higher resolution on relatively short utterances. Fe-CER is compared with other ER estimation models using phonemes, byte-pair encoding tokens as well as words. The performance of the models is measured using normalised root mean square error (nRMSE), which takes into consideration the different distributions of target ERs. Fe-CER trained on Chime5 is shown to outperform the baseline model using word error rate in nRMSE and PCC by 6.00% and 8.79% relative, respectively.**

***Index Terms*—Automatic speech recognition, Error rate estimation, Word error rate, Character error rate, Tokenisation**

I. Introduction

Automatic speech recognition (ASR) output is typically evaluated by measuring an error rate. The errors are determined by comparing ground-truth transcripts, the so-called reference, and the ASR output, referred to as (recognition) hypothesis. This is costly and therefore not available in situations when one just aims to measure the quality in general. An estimate of the confidence of the ASR system on utterances can be used [1]–[5] for this purpose. However, such a method often requires internal information, such as word confidence or acoustic model probabilities, of the ASR system. Access to this information is not always available in commercial systems. To address this issue, methods for estimating the word error rate (WER) of ASR systems' output without reference and the internal information have been investigated [6]–[8]. For example, in [6], WER was estimated using features outside decoding, such as ratio between silences and words. Another series of studies on WER estimation [9]–[12] followed by adopting deep neural networks. The WER estimation models in [11], [12] employed self-supervised representation learning models for extracting audio and text features, which is obtained without any ASR process.

WER estimation based quality estimation was found to have several issues of performance degradation in complex settings. Degradation was observed when the mean and variance of WER on a training dataset were relatively high, especially when the dataset comprises relatively short utterances. For example, when WER is measured of a one word utterance, the quantisation of WER means that it will be 0%, 100%, 200%, etc. The high quantisation noise in this case leads to high means and variances. Inspired by this observation, various tokenisation strategies are employed for error rate (ER) estimation on short utterances in this study. The transcripts are tokenised into either phoneme, character or byte-pair encoding (BPE) tokens. With the tokens, Phoneme Error Rate (PER) [13], [14], Character Error rate (CER) [15]–[17], Token Error Rate (TER) [18], [19] are used for ER estimation for ASR. As transcripts are tokenised in a smaller unit, the denominator of error rate tends to be larger. As a result, the metric becomes higher resolution. This approach is shown to affect the performance of an ER estimation model not only short utterance but also overall ranges. The various tokenisation strategies are evaluated on several ASR corpora: TED-LIUM3(TL3) [20], Chime5 (CH5) [21].

The contributions of this work are:

- an ER estimation model using CER (Fe-CER) for ASR corpus consisting of relatively short utterances
- a comparison of method for ER estimation models using different tokenisation strategies using by normalised root mean square error(nRMSE) as an evaluation metric
- experimental results on a range of corpora.

II. Related Works

Initial work was done by [6] that showed a method using ASR system agnostic features for reference-free ASR error rate estimation. For example, a training instance consists of an utterance, a transcript and WER, while the test instance an utterance and a transcript. In [11], a WER estimation model for multilingual data was proposed, using self-supervised learning representation (SSLR) models. Features for both speech and text were extracted with SSLR models, such as XLSR-53 [22] and XLM-R [23], respectively. The model aggregated features for speech and text by bidirectional long short-term memory and average pooling, respectively. The performance of WER estimation has been evaluated in terms of root mean square errors (RMSE) and Pearson correlation coefficient (PCC). Following the previous studies, a Fast WER estimator (Fe-WER) using average pooling over both speech and text fea-

tures has been proposed. The features were extracted using HuBERT [23] and XLM-R, respectively. The results showed that the computational efficiency for WER estimation has been improved without a performance degradation.

Fe-WER, as illustrated in Fig. 1, is based on a dual-tower architecture [24], [25]. The proposed model comprises two aggregators, one for representing the speech signal and one for text. This is combined with neural networks that output the WER estimation. These aggregators transform the features extracted using SSLR models into a sequence-level representation by averaging frame features over time. The sequence-level representations are combined and taken into multi-layer perceptrons (MLPs) consisting of fully connected layers with rectified linear units (ReLU). The resulting output undergoes a sigmoid function transformation.

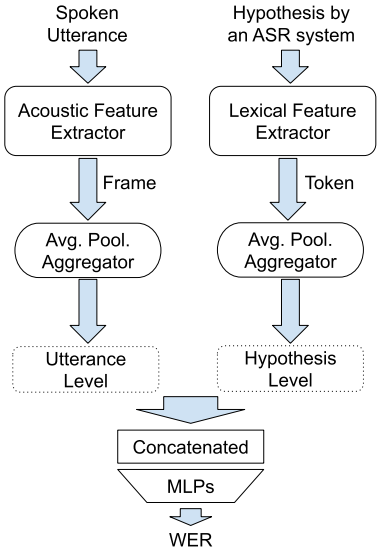


Fig. 1: Architecture of Fe-WER

III. Error rate estimation models

A. Tokenisation

WER is widely used as a target for error rate estimation for ASR, providing an analysis of word-level tokenisation. To address its low resolution on short utterances, other tokenisation schemes such as CER/PER/TER are introduced for error rate estimation. CER offers a finer grained analysis by assessing error rates at the character level instead of word level. Phonemes are determined by mapping words to phonemes, which are the distinct units of sound in a language. Lastly, TER is calculated using byte-pair encoding tokens in the transcripts. An example of tokenisation is shown in Table I. A normalised sentence, "i started reading about alzheimer is and tried to familiarise myself with the research", is assumed before tokenisation.

The target error rates are calculated by first performing dynamic programming alignment, and then using the equation

TABLE I: Example of Tokenisation

	tokenisation
WER	['i', 'started', 'reading', 'about', 'alzheimer', 'is', 'and', 'tried', 'to', 'familiarize', 'myself', 'with', 'the', 'research']
CER	['i', ' ', 's', 't', 'a', 'r', 't', 'e', 'd', ' ', 'r', 'e', 'a', 'd', 'i', 'n', 'g', ' ', 'a', 'b', 'o', 'u', 't', ' ', 'a', 'l', 'z', 'h', 'e', 'i', 'm', 'e', 'r', ' ', 'i', 's', ' ', 'a', 'n', 'd', ' ', 't', 'r', 'i', 'e', 'd', ' ', 't', 'o', ' ', 'f', 'a', 'm', 'i', 'l', 'i', 'a', 'r', 'i', 'z', 'e', ' ', 'm', 'y', 's', 'e', 'l', 'f', ' ', 'w', 'i', 't', 'h', ' ', 't', 'h', 'e', ' ', 'r', 'e', 's', 'e', 'a', 'r', 'c', 'h']
PER	['AY', 'S', 'T', 'AA', 'R', 'T', 'AH', 'D', 'R', 'EH', 'D', 'IH', 'NG', 'AH', 'B', 'AW', 'T', 'AE', 'L', 'Z', 'HH', 'AY', 'M', 'ER', 'IH', 'Z', 'AH', 'N', 'D', 'T', 'R', 'AY', 'D', 'T', 'UW', 'F', 'AH', 'M', 'IH', 'L', 'Y', 'ER', 'AY', 'Z', 'M', 'AY', 'S', 'EH', 'L', 'F', 'W', 'IH', 'DH', 'DH', 'AH', 'R', 'IY', 'S', 'ER', 'CH']
TER	[b'i', b' started', b' reading', b' about', b' al', b'zheimer', b' is', b' and', b' tried', b' to', b' familiar', b'ize', b' myself', b' with', b' the', b' research']

$$ER_x = \frac{S_x + D_x + I_x}{N_x} \tag{1}$$

where x is denoted as W(Word), C(Character), P(Phoneme) and T(Token) and N_x is the number of x in a reference, S, D, I are the number of substitutions, deletions and insertions, respectively.

B. Training objective

The objective function for model training employs the mean squared error (MSE) between target ER and ER estimate, where ER represents the error rate between references and hypotheses, and $\widehat{ER}$ is the model's estimation. Here, N denotes the number of instances in the dataset, and i serves as an index for each instance.

$$MSE_x = \frac{\sum_{i=1}^{N}(ER_{x,i} - \widehat{ER}_{x,i})^2}{N} \tag{2}$$

C. Evaluation

The performance of the ER estimation models with WER, CER, PER and TER is evaluated on test datasets by nRMSE and PCC. As described in Section II, ER estimation models have been evaluated in RMSE and PCC. RMSE represents the average difference between targets and estimates. However, when the performance of ER estimation models using different metrics are measured, their performance can not be directly compared using RMSE due to the different distributions of the target values. To take this into consideration, RMSE is normalised by the standard deviation σ of target error rate. The variance of target ER can be regarded as the MSE of a model that always predicts the mean of the target. By normalising RMSE by the standard deviation of the target ER, how much the performance improvement of the presented model is gained from the model predicting the mean can be measured. It would be less than 1 if there is an improvement. The more

performance is improved, the lower nRMSE becomes. The normalised RMSE (nRMSE) is defined as

$$nRMSE_x = \frac{\sqrt{MSE_x}}{\sigma_x} \tag{3}$$

PCC quantifies the degree of the linear association between two variables. It ranges from -1 to 1, where -1 indicates a complete negative linear correlation, 0 implies no correlation, and +1 signifies a full positive correlation.

$$PCC = \frac{\sum_{i=1}^{N}(ER_{x,i} - \mu_{ER_x})(\widehat{ER}_{x,i} - \mu_{\widehat{ER_x}})}{\sqrt{\sum_{i=1}^{N}(ER_{x,i} - \mu_{ER_x})^2 \sum_{i=1}^{N}(\widehat{ER}_{x,i} - \mu_{\widehat{ER_x}})^2}} \tag{4}$$

where μ_{ER_x} is the mean of ER_x.

IV. EXPERIMENTAL SETUP

A. Data

Two datasets are used in the experiments: TL3 [20] and CH5 [21]. TL3 is a dataset containing speech from public talks on various topics, while CH5 is an ASR corpus of daily conversation in home environments and consists of relatively short utterances. The statistics of TL3 and CH5 are summarised in Table II. The average duration of CH5 utterances is lower than that for TL3. The ER estimation models will be evaluated on CH5 to verify whether a model using a different tokenisation strategy outperforms the Fe-WER model when it is trained on the dataset consisting of short utterances. The reference is preprocessed to be lower-case and to remove punctuation. They are transcribed using the Whisper large model [26]. The hypotheses are preprocessed in the same way.

TABLE II: Statistics of Dataset

		# of segments	total dur.(h)	avg. dur.(s)
TL3	train	123255	200.55	5.86
	dev	1034	1.70	5.93
	eval	842	1.41	6.04
CH5	train	70483	35.86	1.83
	dev	6200	4.41	2.56
	eval	9918	5.23	1.90

B. ER estimation using CER/PER/TER

The performance of ER estimation models using CER/PER/TER is compared with Fe-WER, which is mentioned in section II. The WER/CER/PER/TER estimators utilise average pooling over either the frame or token dimension as an aggregation method. It comprises MLPs with two hidden layers and an output layer, along with activation functions applied to the concatenated feature layer. Additionally, batch normalisation is applied to the output of each layer, and dropout is implemented on the hidden layers.

The hyper-parameters for those are selected through grid search. The optimiser, activation function for the hidden and output layers, and hidden layer dimensions are described on Table III. The estimators are trained using a cosine annealing scheduler and early stopping after 40 epochs.

Regarding tokenisation strategies, the hypothesis is split by white space for WER because after text normalisation they comprise only English letters and spaces without punctuation. For CER, spaces are treated as a character as their position affects the pronunciation. The words are mapped on phonemes using the dictionary provided by TL3 corpus[1] for both TL3 and CH5 for PER. While spaces are ignored for tokenisation at the phoneme level, they are included in BPE tokens. For TER, tiktoken[2] has been adopted. This tokeniser was used for Whisper [26] and GPT models [27].

TABLE III: Train model setup parameters

		Learning Rate	Activation Func	Layers
WER	TL3	0.001	Sigmoid	600, 32
	CH5	0.0003	Sigmoid	300, 16
CER	TL3	0.007	Sigmoid	600, 16
	CH5	0.0007	Sigmoid	300, 16
PER	TL3	0.007	Clamp	300, 32
	CH5	0.0003	Sigmoid	600, 32
TER	TL3	0.003	Sigmoid	600, 32
	CH5	0.0007	Sigmoid	300, 16

V. RESULTS

To test whether there is a correlation between WER and CER/PER/TER, PCC was calculated as shown in Table IV. As shown, WER is highly correlated with those metrics, which indicates a strong positive linear relationship between the two variables.

TABLE IV: PCC between WER and CER/PER/TER

	WER/CER	WER/PER	WER/TER
TL3	0.9429	0.9383	0.9704
CH5	0.9598	0.9565	0.9846

The performance of ER estimation models using different metrics were measured by RMSE and PCC. While Fe-WER outperformed the others on TL3 in PCC, the RMSE of Fe-CER was lowest. On CH5, the performance of Fe-CER was the best in both RMSE and PCC. The results are in Table V.

TABLE V: RMSE and PCC of ER estimation models

		Fe-WER	Fe-CER	Fe-PER	Fe-TER
TL3	RMSE	0.0877	**0.0803**	0.0808	0.0986
	PCC	**0.8995**	0.8989	0.8994	0.8796
CH5	RMSE	0.3202	**0.2821**	0.2956	0.3258
	PCC	0.6154	**0.6695**	0.6505	0.6059

The higher RMSE of Fe-WER on TL3 than that of Fe-CER could have been caused by the different distributions of target

[1]https://www.openslr.org/51

[2]https://github.com/openai/tiktoken

ERs. For example, the average and the standard deviation of target WER were 0.1429 and 0.1997 on TL3, while those of target CER were 0.1061 and 0.1816. The average and the standard deviation of target ERs are summarised in Table VI.

TABLE VI: Average and standard deviation of target ERs

		WER	CER	PER	TER
TL3	average	0.1429	0.1061	0.1066	0.1579
	std.dev.	0.1997	0.1816	0.1839	0.2063
CH5	average	0.3665	0.3174	0.3246	0.3789
	std.dev.	0.4036	0.3783	0.3867	0.4067

To compare the ER estimation models using different metrics, the RMSE was normalised by standard deviation as described in Section III-C. In terms of nRMSE, Fe-WER outperformed the other models in contrast to the result on TL3 in Table V. The nRMSEs of the models are exhibited as Table VII.

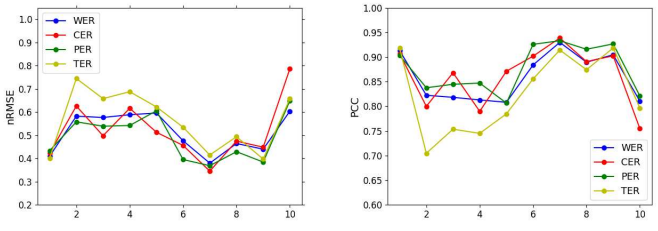
TABLE VII: nRMSE of ER estimation models

		Fe-WER	Fe-CER	Fe-PER	Fe-TER
TL3	nRMSE	**0.4391**	0.4421	0.4394	0.4779
CH5	nRMSE	0.7933	**0.7457**	0.7644	0.8010

The nRMSE of Fe-CER was found to be lower by relative 6% compared to Fe-WER and the PCC of Fe-CER was higher than that of Fe-WER on CH5 by relative 8.79%. On TL3, Fe-WER outperformed the other models in both nRMSE and PCC. Therefore, Fe-CER performed best when it was trained on the relatively short utterances, such as CH5.

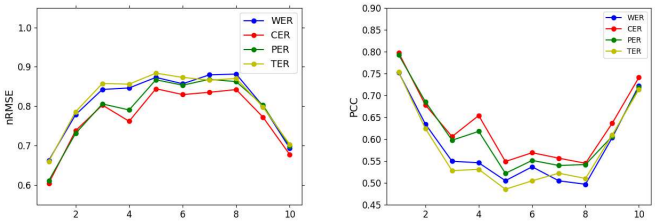
VI. ANALYSIS

The reason different models were used was to see performance improvement on short utterances. For analysis, utterances were sorted by duration in ascending order. Then, they were split into 10 groups. Fig. 2 and 3 display the nRMSE and PCC values for 10 bins of utterances on TL3/CH5 evaluation datasets. While there are relatively small differences between WER and other metrics on the 1st bin of TL3, there are considerable gaps among the metrics on CH5. The ER estimation model using CER showed better performance on CH5. The nRMSE of Fe-CER is lower than that of Fe-WER by relative 8.72%. The PCC of the model is higher than that of Fe-WER by relative 5.96%. Fe-CER outperformed especially in the 4th bin on CH5 in nRMSE and PCC, which are better than those of Fe-WER by relative 9.97% and 19.75%, respectively. The performance of Fe-CER in terms of nRMSE and PCC was better when the model was trained on relatively short utterances such as CH5.



(a) nRMSE (b) PCC

Fig. 2: nRMSE and PCC of ER estimates on TL3. The ordered utterances are grouped into 10 bins.



(a) nRMSE (b) PCC

Fig. 3: nRMSE and PCC of ER estimates on CH5. The ordered utterances are grouped into 10 bins.

VII. CONCLUSION

For evaluation of ASR output on short utterances, an ER estimation model using character error rate is proposed. The performance of ER estimation models with various tokenisation strategies are evaluated using nRMSE and PCC. The result demonstrates that the Fe-CER outperforms the Fe-WER on the CH5 dataset consisting of relatively short utterances by 6.00% and 8.79% relative in both metrics, respectively. In the analysis of performance of the ER estimation models along the utterance duration, Fe-CER consistently outperforms the others on CH5.

ACKNOWLEDGMENT

This work was conducted at the Voicebase/Liveperson Centre of Speech and Language Technology at the University of Sheffield which is supported by Liveperson, Inc..

REFERENCES

[1] A. Kumar, S. Singh, D. Gowda, A. Garg, S. Singh, and C. Kim, "Utterance confidence measure for end-to-end speech recognition with applications to distributed speech recognition scenarios," in *Proc. Interspeech 2020*, 2020, pp. 4357–4361.

[2] A. Woodward, C. Bonnín, I. Masuda, D. Varas, E. Bou-Balust, and J. C. Riveiro, "Confidence Measures in Encoder-Decoder Models for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 611–615.

[3] D. Qiu, Q. Li, Y. He, Y. Zhang, B. Li, L. Cao, R. Prabhavalkar, D. Bhatia, W. Li, K. Hu, T. N. Sainath, and I. McGraw, "Learning word-level confidence for subword end-to-end asr," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6393–6397.

[4] Q. Li, D. Qiu, Y. Zhang, B. Li, Y. He, P. C. Woodland, L. Cao, and T. Strohman, "Confidence estimation for attention-based sequence-to-sequence models for speech recognition," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6388–6392.

[5] D. Oneaţă, A. Caranica, A. Stan, and H. Cucu, "An evaluation of word-level confidence estimation for end-to-end automatic speech recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 258–265.

[6] M. Negri, M. Turchi, J. G. C. de Souza, and D. Falavigna, "Quality estimation for automatic speech recognition," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, J. Tsujii and J. Hajic, Eds., Dublin, Ireland, Aug. 2014, pp. 1813–1823, Dublin City University and Association for Computational Linguistics.

[7] J. G. C. de Souza, H. Zamani, M. Negri, M. Turchi, and D. Falavigna, "Multitask learning for adaptive quality estimation of automatically transcribed utterances," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, R. Mihalcea, J. Chai, and A. Sarkar, Eds., Denver, Colorado, May–June 2015, pp. 714–724, Association for Computational Linguistics.

[8] S. Jalalvand, M. Negri, D. Falavigna, M. Matassoni, and M. Turchi, "Automatic quality estimation for asr system combination," *Comput. Speech Lang.*, vol. 47, no. C, pp. 214–239, jan 2018.

[9] A. Ali and S. Renals, "Word error rate estimation for speech recognition: e-WER," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia, July 2018, pp. 20–24, Association for Computational Linguistics.

[10] A. Ali and S. Renals, "Word error rate estimation without ASR output: e-WER2," in *Proc. Interspeech 2020*, 2020, pp. 616–620.

[11] S. A. Chowdhury and A. Ali, "Multilingual word error rate estimation: e-WER3," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[12] C. Park, C. Lu, M. Chen, and T. Hain, "Fast word error rate estimation using self-supervised representations for speech and text," *arXiv preprint arXiv:2310.08225*, 2023.

[13] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, and F. Metze, "Universal phone recognition with a multilingual allophone system," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8249–8253.

[14] A. Fang, S. Filice, N. Limsopatham, and O. Rokhlenko, "Using phoneme representations to build predictive models robust to asr errors," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 2020, SIGIR '20, p. 699–708, Association for Computing Machinery.

[15] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 798–805.

[16] H. Sato, T. Ochiai, M. Delcroix, K. Kinoshita, N. Kamo, and T. Moriya, "Learning to enhance or not: Neural network-based switching of enhanced and observed signals for overlapping speech recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6287–6291.

[17] S. Zhang, C.-T. Do, R. Doddipatla, and S. Renals, "Learning noise invariant features through transfer learning for robust end-to-end speech recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7024–7028.

[18] S.-P. Chuang, H.-J. Chang, S.-F. Huang, and H.-y. Lee, "Non-autoregressive mandarin-english code-switching speech recognition," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 465–472.

[19] X. Zhou, E. Yılmaz, Y. Long, Y. Li, and H. Li, "Multi-Encoder-Decoder Transformer for Code-Switching Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 1042–1046.

[20] 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," in *Speech and Computer*, A. Karpov, O. Jokisch, and R. Potapova, Eds., Cham, 2018, pp. 198–208, Springer International Publishing.

[21] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *Proc. Interspeech 2018*, 2018, pp. 1561–1565.

[22] S. Ruder, A. Søgaard, and I. Vulić, "Unsupervised cross-lingual representation learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, P. Nakov and A. Palmer, Eds., Florence, Italy, July 2019, pp. 31–38, Association for Computational Linguistics.

[23] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 8440–8451, Association for Computational Linguistics.

[24] J. Yang et al., "Mixed negative sampling for learning two-tower neural networks in recommendations," in *Companion Proceedings of the Web Conference 2020*, New York, NY, USA, 2020, WWW '20, p. 441–447, Association for Computing Machinery.

[25] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, New York, NY, USA, 2013, CIKM '13, p. 2333–2338, Association for Computing Machinery.

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*. 2023, ICML'23, JMLR.org.

[27] OpenAI, S. Adler, Agarwal, et al., "Gpt-4 technical report," *arXiv.org*, 2023.