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

Residual Energy-Based Models for End-to-End Speech Recognition

Qiujia Li[1*], *Yu Zhang*[2], *Bo Li*[2], *Liangliang Cao*[2], *Philip C. Woodland*[1]
[1]University of Cambridge, UK
[2]Google LLC, USA
{ql264,pcw}@eng.cam.ac.uk, {ngyuzh,boboli,llcao}@google.com

Abstract

End-to-end models with auto-regressive decoders have shown impressive results for automatic speech recognition (ASR). These models formulate the sequence-level probability as a product of the conditional probabilities of all individual tokens given their histories. However, the performance of locally normalised models can be sub-optimal because of factors such as exposure bias. Consequently, the model distribution differs from the underlying data distribution. In this paper, the residual energy-based model (R-EBM) is proposed to complement the auto-regressive ASR model to close the gap between the two distributions. Meanwhile, R-EBMs can also be regarded as utterance-level confidence estimators, which may benefit many downstream tasks. Experiments on a 100hr LibriSpeech dataset show that R-EBMs can reduce the word error rates (WERs) by 8.2%/6.7% while improving areas under precision-recall curves of confidence scores by 12.6%/28.4% on test-clean/test-other sets. Furthermore, on a state-of-the-art model using self-supervised learning (wav2vec 2.0), R-EBMs still significantly improves both the WER and confidence estimation performance.

Index Terms: energy-based model, end-to-end speech recognition, confidence estimation.

1. Introduction

End-to-end trainable speech recognition models have shown promising performance in automatic speech recognition (ASR) [1, 2, 3, 4, 5], especially using recurrent neural network transducers (RNN-Ts) [6] and attention-based encoder-decoder models [7]. One common characteristic of these two types of model is that the decoder learns the conditional distribution of the current output token given all the history tokens. This leads to *locally normalised* auto-regressive models where the output probability distributions are normalised per output token. The final probability of the output sequence is obtained by computing the product of a series of conditional probabilities. This allows end-to-end models to be trained efficiently using the maximum likelihood criterion and yields good ASR performance.

However, locally normalised models can in practice be sub-optimal. First, history tokens are the ground truth during training while history tokens may contain errors during inference as they are generated sequentially. This training and inference mismatch is referred to as exposure bias [8]. To this end, scheduled sampling has been proposed to allow generated tokens to appear in the history during training with a certain probability according to a specific schedule [9]. Using a larger beam during decoding or incorporating beam search heuristics can also help reduce search errors [10]. Secondly, the locally normalised model trained with maximum likelihood may not be optimal in terms of the final evaluation criterion, *i.e.* word error rate (WER) for ASR. Sequence-level training criteria that directly minimise the number of word errors have been proposed to address this mismatch [11]. Although all of the above techniques can improve model performance to some extent, the model distribution is still based on the product of locally normalised distributions, which may differ from the data distribution [12].

Residual energy-based models (R-EBMs), which were first used for text generation [12], use energy-based models (EBMs) [13] to learn from the residual errors of an auto-regressive generator to reduce the gap between the model and data distributions. An R-EBM can also be viewed as a discriminator between generated samples from an auto-regressive model and real data samples [12]. For end-to-end ASR, the R-EBM, conditioned on acoustic features, aims to distinguish the model-generated hypotheses from the ground truth transcriptions. R-EBMs are trained using conditional noise contrastive estimation (NCE) [14, 15]. For a given utterance, the positive sample is the ground truth whereas the n-best hypotheses are taken as the negative samples. During inference, the auto-regressive model first generates a list of hypotheses for each utterance, and the best hypothesis from the joint model can be obtained by using the combined score of the log-likelihood from the auto-regressive model and the negative energy value from the R-EBM.

As a discriminator between correct and erroneous hypotheses, R-EBMs can also produce *utterance-level confidence scores* for end-to-end ASR models. Some downstream tasks only require utterance-level confidence scores, such as data selection for semi-supervised learning [16, 17], and hypothesis-level model combination [18]. Compared to token or word level confidence, direct modelling of utterance-level confidence scores implicitly takes deletion errors into account, and does not require calibration of the confidence scores (e.g. using piece-wise linear mapping [19]) before taking the average for utterance-level scores. Previously, many model-based methods have been used for confidence estimation for both conventional [20, 21, 22, 23] and end-to-end ASR [24, 18, 25, 26, 27, 28], and n-best re-ranking models [29, 30, 31, 32] have been proposed to improve WER. The R-EBM is a single model that can improve speech recognition performance and utterance-level confidence estimation performance at the same time.

In the rest of this paper, Section 2 formulates the R-EBM for end-to-end ASR. Then Sec. 3 describes the data and models used for the experiments. Experimental results are presented in Sec. 4. Further analysis on the behaviour of R-EBMs is given in Sec. 5, followed by conclusions in Sec. 6.

2. Residual Energy-Based Models

2.1. Overview

ASR models the conditional distribution of the text sequence $\mathbf{y}$ given the input acoustic sequence $\mathbf{X}$. For two state-of-the-art types of end-to-end trainable systems, namely attention-based sequence-to-sequence models and recurrent neural network transducers (RNN-Ts), the model distribution can be ex-

*Work was done while the author interned at Google.

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-690

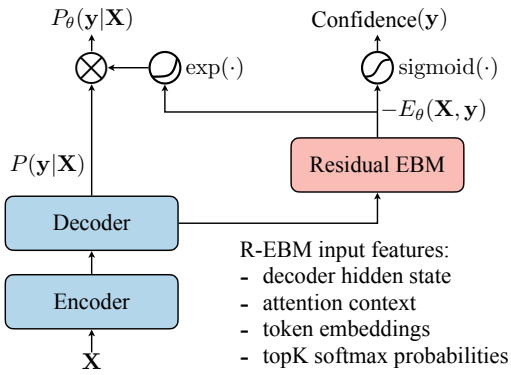

Figure 1: *Schematic of a residual energy-based model (R-EBM) for an end-to-end ASR model. The baseline auto-regressive model is fixed during R-EBM training.*

panded using chain rule as in Eqn. (1).

$$P(\mathbf{y}|\mathbf{X}) = P(y_1|\mathbf{X}) \prod_{t=2}^{T} P(y_t|y_{1:t-1}, \mathbf{X}). \tag{1}$$

These end-to-end models are locally normalised as an auto-regressive decoder predicts the next token based on the acoustic features and the past tokens for each step. The tokens used for end-to-end ASR models can be a set of characters, word pieces or even words. Given an unlimited model capacity, the auto-regressive model trained using maximum likelihood has the potential to learn the true data distribution perfectly [12]. However, in practice, there are various drawbacks associated with locally normalised models. Specifically for end-to-end ASR, exposure bias [8] can be an issue during decoding when errors exist in the history; more search errors are likely to occur since only top few hypotheses are kept; and token-level maximum likelihood training does not directly minimise word error rates [11]. Therefore, the distribution learned by a locally normalised auto-regressive model $P(\mathbf{y}|\mathbf{X})$ may not match the real data distribution P_{data}. To have a globally normalised model, [12] proposed residual energy-based models (R-EBM) that learns "residual" errors of the auto-regressive model to better match the data distribution. The R-EBM, parameterised by $\boldsymbol{\theta}$, can be formulated as

$$P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{X}) = \frac{P(\mathbf{y}|\mathbf{X}) \exp\big(-E_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{y})\big)}{Z_{\boldsymbol{\theta}}(\mathbf{X})}, \tag{2}$$

where $P_{\boldsymbol{\theta}}$ is the joint model, $E_{\boldsymbol{\theta}}$ is the residual energy function, and $Z_{\boldsymbol{\theta}}$ is the partition function for the energy-based model, which can be computed as

$$Z_{\boldsymbol{\theta}}(\mathbf{X}) = \sum_{y} P(\mathbf{y}|\mathbf{X}) \exp\big(-E_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{y})\big). \tag{3}$$

In principle, the R-EBM itself can be any model that takes a pair of acoustic sequence $\mathbf{X}$ and hypothesis sequence $\mathbf{y}$ to produce a scalar value $(-E_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{y}))$. In this paper, the R-EBM is a sequence model that takes features including the current decoder hidden state, the acoustic context vector from the attention mechanism, the output token embeddings, and the topK softmax probabilities for each output step [24]. Then the hidden representations for each output token are gathered by mean pooling before being passed to the output layer that uses sigmoid activation function. Since R-EBMs operate at the sequence-level, bi-directional models can be used.

2.2. Training

However, for a system with a token vocabulary size V and a maximum output sequence length T, the partition function quickly becomes intractable as the summation is over V^T possible sequences. With the baseline auto-regressive model P fixed, conditional noise contrastive estimation (NCE) [14] can be used to train the R-EBM where the noise distribution is the auto-regressive ASR model P. The loss can be expressed as

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{X}} \Big\{ \mathbb{E}_{\mathbf{y}^+ \sim P_{\text{data}}(\cdot|\mathbf{X})} \log \frac{1}{1+\exp\big(E_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{y}^+)\big)} + \mathbb{E}_{\mathbf{y}^- \sim P(\cdot|\mathbf{X})} \log \frac{1}{1+\exp\big(-E_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{y}^-)\big)} \Big\}, \tag{4}$$

where $\mathbf{y}^+$ are samples from the data distribution and $\mathbf{y}^-$ are samples from the noise distribution. For ASR, the noise samples are the n-best hypotheses from the auto-regressive model via beam search, and the positive samples are the corresponding ground truth transcription. Thus, the R-EBM is effectively a discriminator between incorrect sequences Y^- and correct sequences Y^+, trained using the binary cross-entropy loss,

$$\mathcal{L}(\mathbf{X};\boldsymbol{\theta}) \approx \frac{1}{|Y^+|} \sum_{\mathbf{y}\in Y^+} \log \frac{1}{1+\exp\big(E_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{y})\big)} + \frac{1}{|Y^-|} \sum_{\mathbf{y}\in Y^-} \log \frac{1}{1+\exp\big(-E_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{y})\big)}, \tag{5}$$

where $Y^+ \cup Y^- = \{\mathbf{y}_{\text{ground truth}}, \textsc{BeamSearch}(\mathbf{X}, n)\}$. Note that there may be more than one element in Y^+ since multiple tokenization using sub-word units for the ground truth may exist or the ground truth is among the n-best hypotheses. Finally, the parameters for the R-EBM can be estimated by optimising a binary classifier over all the utterances in the entire dataset.

2.3. Inference

For a given utterance during inference, the log-likelihood of a hypothesis and the negative energy value of the hypothesis are added to obtain the joint score as in Eqn. (6).

$$\log P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{X}) \propto \log P(\mathbf{y}|\mathbf{X}) - E_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{y}). \tag{6}$$

With a shared partition function, the n-best hypotheses can be re-ranked based on the joint scores to yield the best candidate.

From another perspective, the R-EBM is a binary classifier that learns to assign scores close to 1 for correct hypotheses and 0 for erroneous ones, which is also the objective for utterance-level confidence scores [26]. Confidence scores can be used to automatically assess the quality of transcriptions of ASR systems. For applications where utterance-level confidence scores are also required, the R-EBM can be used to achieve two aims. The pre-sigmoid values of R-EBM $(-E_{\boldsymbol{\theta}}(\mathbf{X},\mathbf{y}))$ can be used to rerank n-best hypotheses to lower the word error rate of the ASR system while the post-sigmoid values can be used as a model-based confidence measure.

3. Experimental Setup

3.1. Data

For training the baseline ASR model and R-EBMs, the "train-clean-100" subset of the LibriSpeech corpus [33] was used. The

100-hour training set has 28.5k utterances and the average duration per utterance is 12.7 seconds. The dev and test sets are dev-clean/dev-other and test-clean/test-other with each one being over 5 hours. This dataset contains read speech from audiobooks. The text data for the language model (LM) is the LibriSpeech LM corpus with 40 million sentences. The acoustic features are 80-dimensional filterbank coefficients with Δ and $\Delta\Delta$ and the modelling units are a set of 1024 word-pieces [34] derived from the LibriSpeech 100h training transcriptions.

3.2. Models

The baseline model architecture is an attention-based sequence-to-sequence model, where the encoder has 2 convolutional layers with a stride of 2 followed by a 4-layer bi-directional long short-term memory (LSTM) network with 1024 units in each direction. The decoder has a 2-layer uni-directional LSTM network with 1024 units. The total number of parameters is 145 million. The Adam optimiser is used with a learning rate of 0.001 and batch size of 512. Training techniques such as SpecAugment [35], dropout, label smoothing, Gaussian weight noise and exponential moving average are used to improve performance. The LM has 2 uni-directional LSTM layers with 1024 units in each layer. Shallow fusion [36] is used for decoding and for generating n-best hypotheses for R-EBMs. Hyperparameters for the attention-based models, the language model and beam search are tuned on the dev sets.

The R-EBMs are trained with the baseline ASR model fixed. The n-best hypotheses of the training set are generated on-the-fly with beam size n and random SpecAugment masks. For time masks, instead of 2 masks with a maximum of 40 frames per mask for the baseline ASR model, 10 masks with a maximum of 50 frames per mask are used for R-EBM training. This is to simulate the errors made by the model during inference and the randomness of masks allows diverse errors to appear during training. The WER on the augmented training set should ideally match that of the dev set.

4. Experimental Results

4.1. Length Normalisation and Log-linear Interpolation

Table 1: *Impact on WERs (%) by applying length normalisation (LN) for log-likelihood score on dev-clean/dev-other sets. The R-EBM is a uni-directional LSTM and the beam size is 8.*

	1-best	R-EBM	joint
w/o LN	5.84/18.48	5.44/18.13	5.38/17.91
w/ LN	5.28/18.41		5.09/17.77

After beam search, n-best hypotheses are determined by keeping the n terminated hypotheses with the highest sequence-level log-likelihood. However, this criterion may favour shorter hypotheses when finding the 1-best. Therefore, normalising the log-likelihood by the number of tokens in each hypothesis results in a slightly lower WER as shown in Table 1. Length normalisation (LN) becomes more important when the n is large. When combining the log-likelihood score $\log P(\mathbf{y}|\mathbf{X})$ with the negative energy score $-E_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{y})$ for the joint score, a coefficient α is tuned on dev sets to minimise WER and accommodate potentially different numerical ranges as in Equation (7).

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \ \frac{\log P(\mathbf{y}|\mathbf{X})}{|\mathbf{y}|} - \alpha E_{\boldsymbol{\theta}}(\mathbf{X}, \mathbf{y}). \tag{7}$$

Table 1 shows that ranking the n-best hypotheses just using the R-EBM scores reduces the WER compared to 1-best WER without LN. After log-linear interpolation, the WER of the joint model is lower than the 1-best results with or without LN. Therefore, all the following experiments will use LN.

4.2. R-EBM Architecture

Table 2: *Comparison of WERs (%) by using uni-directional and bi-directional LSTMs for R-EBMs on dev-clean/dev-other sets. The beam size for decoding is 8.*

	1-best	R-EBM	joint
uni-directional	5.28/18.41	5.44/18.13	5.09/17.77
bi-directional		5.33/18.05	5.07/17.69

As a globally normalised model, an R-EBM can be bi-directional and take advantage of the full context in the hypotheses. Table 2 compares the performance of uni-directional and bi-directional R-EBMs. Both R-EBMs have two layers of LSTMs with 512 units in each direction. Since the bi-directional model performs slightly better, all the following experiments will use bi-directional LSTMs for R-EBMs.

4.3. Effect of the Size of N-best Lists

Table 3: *WERs of the ASR model and joint models with various numbers of hypotheses used for training and inference.*

	n	oracle	1-best	R-EBM	joint	WERR
dev-clean	4	4.19	5.41	5.58	5.30	2.0
	8	3.52	5.28	5.33	5.07	4.0
	16	3.03	5.17	5.23	4.86	6.0
	32	2.68	5.23	5.33	4.80	8.2
dev-other	4	17.10	19.04	19.07	18.71	1.7
	8	15.33	18.41	18.05	17.69	3.9
	16	13.96	18.06	17.67	17.12	5.2
	32	12.93	17.98	17.31	16.75	6.8

The performance ceiling of the joint model is the oracle WER of the n-best hypotheses. In this section, the lengths of n-best lists range from 4 to 32 for both training and inference. Table 3 shows that oracle WERs improve consistently when larger n-best lists are used whereas 1-best WERs only has minor reduction. With more hypotheses available, the joint WERs reduces significantly together with WERs ranked by R-EBM scores only. The last column in Table 3 shows that WER relative reduction (WERR) of the joint model over 1-best steadily increases with larger n-best lists, which indicates that the gain from the R-EBMs outpaces that of the 1-best. With 32-best, 8.2%/6.8% WERRs are obtained for dev-clean/dev-other sets. Top 32-best will be used for all the following experiments.

4.4. Recognition and Confidence Estimation Performance

Based on previous experimental results on dev sets, bi-directional LSTMs with 32-best hypotheses and length normalisation are used as the best setup. By applying the log-linear combination coefficients α tuned on dev-clean/dev-other sets, the results on two test sets are shown in Table 4. The area under the precision-recall curve (AUC) is used as the metric for confidence estimation [23, 24]. AUC is a number between 0 and 1 and higher values indicate more reliable confidence

Table 4: *Recognition and confidence performance on LibriSpeech test sets. The average of token-level scores from CEM is used as utterance-level scores. For both CEM and R-EBM, the best interpolation coefficients are tuned on the dev sets.*

	test-clean		test-other	
	WER ↓	AUC ↑	WER ↓	AUC ↑
baseline	5.61	0.684	18.68	0.529
+ CEM [24]	5.59	0.697	18.44	0.501
+ R-EBM	**5.15**	**0.770**	**17.42**	**0.679**

scores. Also in Table 4, the second row corresponds to the confidence estimation module (CEM) [24], which predicts a confidence score for each token in the hypothesis sequence using the same input features as R-EBMs. By averaging the token-level scores[1], utterance-level scores can be used to combine with the baseline model. Note that although the CEM yields improved confidence at token and word levels as in [24], the utterance-level confidence may under-perform the baseline log-likelihood score. Since the R-EBM is directly optimised for the utterance-level confidence, issues such as multiple tokenisations for the same word or sequence and deletion errors are addressed implicitly during training. As a result, the R-EBM reduces WERs and significantly improves utterance-level confidence.

4.5. Scalability

Table 5: *Recognition and confidence performance on LibriSpeech test sets when the encoder is initialised using a pre-trained wav2vec 2.0 (w2v2) as a stronger baseline.*

	test-clean		test-other	
	WER ↓	AUC ↑	WER ↓	AUC ↑
w2v2	2.63	0.786	4.74	0.684
+ R-EBM	**2.49**	**0.928**	**4.53**	**0.890**

This section investigates the situation when the auto-regressive ASR model has seen much more data such that the baseline WER is far lower. In this set of experiments, the encoder of the ASR model is first initialised with the pre-trained wav2vec 2.0 (w2v2) model [37] trained on 57.7 thousand hours of unlabelled speech data from Libri-light [38] and then fine-tuned on the same amount of labelled data ("train-clean-100") as before[2]. Although the WERs in Table 5 are much lower than in Table 4, the joint model further yields 5.3%/4.4% WERRs on test-clean/test-other. Meanwhile, AUCs are much higher with more unlabelled data as expected, but R-EBMs can furthur boost confidence estimation significantly.

5. Analysis

5.1. Relative Improvement by Utterance Length

Figure 2 shows the breakdown of WERR for the joint model and the oracle hypotheses with respect to the number of words in reference sequences. Oracle WERR is lower for longer utterances as the number of alternatives per word is fewer with

[1]For each hypothesis, the mean of the pre-sigmoid logits of all tokens is used for n-best reranking, whereas the mean of the post-sigmoid confidence scores of all tokens is used for confidence evaluation.

[2]Out implementation follows [17], which shows the state-of-the-art performance on LibriSpeech.

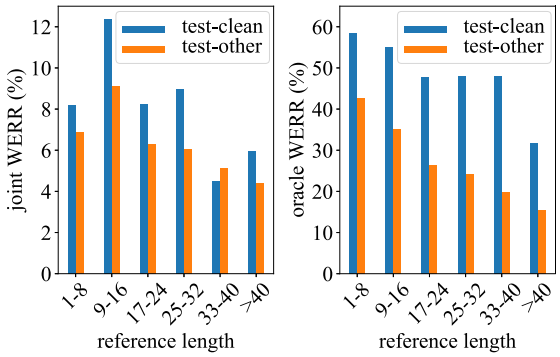

Figure 2: *WERR with respect to the number of words in reference sequences. Left is the WERR of joint over 1-best and right is the WERR of oracle over 1-best. 32-best hypotheses are used.*

a given number of top hypotheses. The general trend of the joint WERR follows the trend of the oracle WERR except for short utterances (1-8 words in reference). We hypothesise that R-EBMs may need more global context information to give a higher WER reduction.

5.2. Distribution Matching

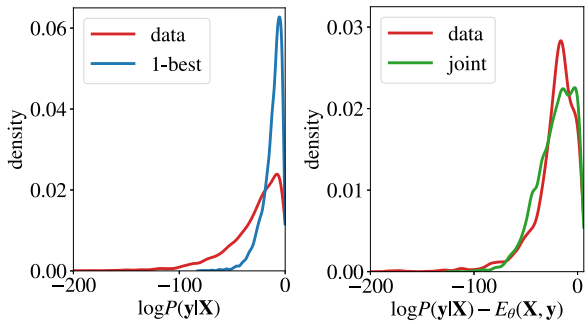

Figure 3: *Density plot of log-probability scores using the baseline model (left) and the joint model (right) on test-other set.*

If the joint model $P_{\boldsymbol{\theta}}$ matches the data distribution P_{data} better, then statistics computed on a large set of samples from the two distributions should also match [37]. Figure 3 shows the density plot of the log-likelihood scores (left) and the joint model scores (right) on test-other set. Red lines correspond to the score distributions of the ground truth transcriptions. The distribution of log-likelihood scores of the best hypotheses from the auto-regressive model does not match the data distribution well. However, the distribution from the joint model is much closer to the data distribution.

6. Conclusions

This paper proposes to use residual energy-based models (R-EBMs) to complement locally normalised auto-regressive end-to-end ASR models. R-EBM is globally normalised as it learns from the residual error of the locally normalised model. R-EBMs can also be viewed as an utterance-level confidence estimator for ASR. Experiments show that R-EBMs can reduce speech recognition error rates while improving the confidence scores at the utterance level, even on top of a state-of-the-art baseline model trained using wav2vec 2.0. Further analysis shows that the performance of R-EBM may depend on the amount of context, and confirms that the R-EBM closes the gap between the model distribution and the data distribution.

7. References

[1] C.-C. Chiu, T. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," *Proc. ICASSP*, Calgary, 2018.

[2] Z. Tüske, G. Saon, K. Audhkhasi, and B. Kingsbury, "Single headed attention based sequence-to-sequence model for state-of-the-art results on Switchboard-300," *Proc. Interspeech*, Shanghai, 2020.

[3] T. Sainath, Y. He, B. Li, A. Narayanan, R. Pang, A. Bruguier, S.-Y. Chang, W. Li, R. Alvarez, Z. Chen, C.-C. Chiu, D. García, A. Gruenstein, K. Hu, M. Jin, A. Kannan, Q. Liang, I. Mcgraw, C. Peyser, R. Prabhavalkar, G. Pundak, D. Rybach, Y. Shangguan, Y. Sheth, T. Strohman, M. Visontai, Y. Wu, Y. Zhang, and D. Zhao, "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," *Proc. ICASSP*, Barcelona, 2020.

[4] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, and Y. Zhang, "Recent developments on ESPnet toolkit boosted by conformer," *Proc. ICASSP*, Toronto, 2021.

[5] J. Li, R. Zhao, Z. Meng, Y. Liu, W. Wei, S. Parthasarathy, V. Mazalov, Z. Wang, L. He, S. Zhao, and Y. Gong, "Developing RNN-T models surpassing high-performance hybrid models with customization capability," *Proc. Interspeech*, Shanghai, 2020.

[6] A. Graves, "Sequence transduction with recurrent neural networks," *Proc. ICML Workshop on Representation Learning*, Edinburgh, 2012.

[7] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Proc. NIPS*, Montreal, 2015.

[8] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *Proc. ICLR*, San Juan, 2016.

[9] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," *Proc. NIPS*, Montreal, 2015.

[10] J. Chorowski and N. Jaitly, "Towards better decoding and language model integration in sequence to sequence models," *Proc. Interspeech*, Stockholm, 2017.

[11] R. Prabhavalkar, T. Sainath, Y. Wu, P. Nguyen, Z. Chen, C.-C. Chiu, and A. Kannan, "Minimum word error rate training for attention-based sequence-to-sequence models," *Proc. ICASSP*, Calgary, 2018.

[12] Y. Deng, A. Bakhtin, M. Ott, A. Szlam, and M. Ranzato, "Residual energy-based models for text generation," *Proc. ICLR*, Addis Ababa, 2020.

[13] Y. LeCun, S. Chopra, R. Hadsell, A. Ranzato, and F. Huang, "A tutorial on energy-based learning," *Predicting Structured Data*, 2006.

[14] Z. Ma and M. Collins, "Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency," *Proc. EMNLP*, Brussels, 2018.

[15] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," *Proc. AISTATS*, Sardinia, 2010.

[16] D. 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," *Proc. Interspeech*, Shanghai, 2020.

[17] Y. Zhang, J. Qin, D. 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," *Proc. NeurIPS SAS Workshop*, Vancouver, 2020.

[18] D. Qiu, Q. Li, Y. He, Y. Zhang, B. Li, L. Cao, R. Prabhavalkar, D. Bhatia, W. Li, K. Hu, T. Sainath, and I. Mcgraw, "Learning word-level confidence for subword end-to-end ASR," *Proc. ICASSP*, Toronto, 2021.

[19] G. Evermann and P. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," *Proc. ICASSP*, Istanbul, 2000.

[20] K. Kalgaonkar, C. Liu, Y. Gong, and K. Yao, "Estimating confidence scores on ASR results using recurrent neural networks," *Proc. ICASSP*, Brisbane, 2015.

[21] M. Á. Del-Agua, A. Giménez, A. Sanchís, J. C. Saiz, and A. Juan, "Speaker-adapted confidence measures for ASR using deep bidirectional recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2018.

[22] Q. Li, P. Ness, A. Ragni, and M. Gales, "Bi-directional lattice recurrent neural networks for confidence estimation," *Proc. ICASSP*, Brighton, 2019.

[23] A. Ragni, Q. Li, M. Gales, and Y. Wang, "Confidence estimation and deletion prediction using bidirectional recurrent neural networks," *Proc. SLT*, Athens, 2018.

[24] 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," *Proc. ICASSP*, Toronto, 2021.

[25] D. Oneata, A. Caranica, A. Stan, and H. Cucu, "An evaluation of word-level confidence estimation for end-to-end automatic speech recognition," *Proc. SLT*, Shenzhen, 2021.

[26] 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," *Proc. Interspeech*, Shanghai, 2020.

[27] A. Woodward, C. Bonnín, I. Masuda, D. Varas, E. Bou, and J. C. Riveiro, "Confidence measures in encoder-decoder models for speech recognition," *Proc. Interspeech*, Shanghai, 2020.

[28] D. Qiu, Y. He, Q. Li, Y. Zhang, L. Cao, and I. McGraw, "Multi-task learning for end-to-end ASR word and utterance confidence with deletion prediction," *Proc. Interspeech*, Brno, 2021.

[29] A. Ogawa, M. Delcroix, S. Karita, and T. Nakatani, "Rescoring n-best speech recognition list based on one-on-one hypothesis comparison using encoder-classifier model," *Proc. ICASSP*, Calgary, 2018.

[30] Q. Li, C. Zhang, and P. Woodland, "Integrating source-channel and attention-based sequence-to-sequence models for speech recognition," *Proc. ASRU*, Singapore, 2019.

[31] T. Sainath, R. Pang, D. Rybach, Y. He, R. Prabhavalkar, W. Li, M. Visontai, Q. Liang, T. Strohman, Y. Wu, I. Mcgraw, and C.-C. Chiu, "Two-pass end-to-end speech recognition," *Proc. Interspeech*, Graz, 2019.

[32] E. Variani, T. Chen, J. Apfel, B. Ramabhadran, S. Lee, and P. Moreno, "Neural oracle search on N-best hypotheses," *Proc. ICASSP*, Barcelona, 2020.

[33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *Proc. ICASSP*, Brisbane, 2015.

[34] M. Schuster and K. Nakajima, "Japanese and Korean voice search," *Proc. ICASSP*, Kyoto, 2012.

[35] D. 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," *Proc. Interspeech*, Graz, 2019.

[36] Çaglar Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," *ArXiv.org*, 1503.03535, 2015.

[37] A. Baevski, H. Zhou, A. rahman Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. NeurIPS*, Vancouver, 2020.

[38] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. rahman Mohamed, and E. Dupoux, "Libri-light: A benchmark for ASR with limited or no supervision," *Proc. ICASSP*, Barcelona, 2020.