

# CONFIDENCE ESTIMATION FOR ATTENTION-BASED SEQUENCE-TO-SEQUENCE MODELS FOR SPEECH RECOGNITION

Qiujia Li<sup>1\*</sup>, David Qiu<sup>2</sup>, Yu Zhang<sup>2</sup>, Bo Li<sup>2</sup>, Yanzhang He<sup>2</sup>,  
Philip C. Woodland<sup>1</sup>, Liangliang Cao<sup>2</sup>, Trevor Strohman<sup>2</sup>

<sup>1</sup> University of Cambridge, UK, <sup>2</sup> Google LLC, USA

<sup>1</sup>{q1264,pcw}@eng.cam.ac.uk, <sup>2</sup>{qdavid,ngyuzh,boboli,yanzhanghe,llcao,strohman}@google.com

## ABSTRACT

For various speech-related tasks, confidence scores from a speech recogniser are a useful measure to assess the quality of transcriptions. In traditional hidden Markov model-based automatic speech recognition (ASR) systems, confidence scores can be reliably obtained from word posteriors in decoding lattices. However, for an ASR system with an auto-regressive decoder, such as an attention-based sequence-to-sequence model, computing word posteriors is difficult. An obvious alternative is to use the decoder softmax probability as the model confidence. In this paper, we first examine how some commonly used regularisation methods influence the softmax-based confidence scores and study the overconfident behaviour of end-to-end models. Then we propose a lightweight and effective approach named confidence estimation module (CEM) on top of an existing end-to-end ASR model. Experiments on LibriSpeech show that CEM can mitigate the overconfidence problem and can produce more reliable confidence scores with and without shallow fusion of a language model. Further analysis shows that CEM generalises well to speech from a moderately mismatched domain and can potentially improve downstream tasks such as semi-supervised learning.

**Index Terms**— confidence scores, end-to-end ASR

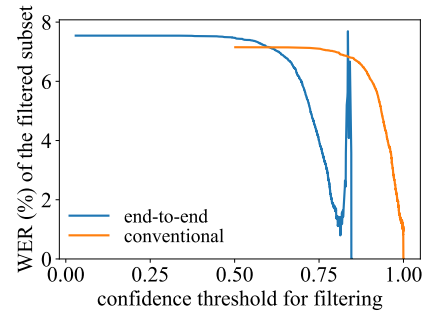
## 1. INTRODUCTION

Confidence scores have been an intrinsic part of automatic speech recognition (ASR) systems [1–3]. Many speech-related applications depend on high-quality confidence scores to mitigate error from speech recognisers. For example, in semi-supervised learning and active learning, utterances with highly confident hypotheses are selected to further improve ASR performance [4–6]. Confidence scores are also used in dialogue systems where queries with low confidence may be returned to users for clarification [5]. As an indication of ASR uncertainty, confidence scores can play an role in speaker adaptation [7], and system combination [8].

In conventional HMM-based systems, reliable confidence scores can be easily obtained by computing word posterior probabilities from compact representations of the hypotheses space, *e.g.* lattices or confusion networks [8, 9]. Improved confidence estimation can be achieved by using model-based approaches, such as conditional random fields [10], recurrent neural networks [11, 12] and graph neural networks [13], or leveraging more related information including phonetics, word/phone duration and language models [2, 14].

Recently, end-to-end speech recognition has achieved promising performance over the conventional systems [15]. “End-to-end

systems” refers to end-to-end differentiable and trainable neural networks, in contrast to modular systems with separate acoustic models, language models and token passing decoders. As end-to-end speech recognition has various modelling and engineering advantages, they are becoming more widely adopted [16]. One class of the end-to-end systems is attention-based sequence-to-sequence models [17, 18]. The auto-regressive decoder in end-to-end systems implies that compact representations like lattices cannot be constructed for attention-based models, except using specific decoder architectures and heuristics [19]. Consequently, computing word posteriors in the end-to-end hypothesis space becomes prohibitively expensive. A greedy approximation would be taking the softmax probability from each step of the decoder as the confidence scores for each token [20]. However, the quality of confidence estimation by softmax probabilities may be very poor [21].



**Fig. 1:** Filtering behaviour of a conventional HMM-based system and an attention-based sequence-to-sequence model based on confidence scores. Utterances with confidence higher than a threshold (x-axis) are selected, and word error rate (WER) of the filtered subset (y-axis) are plotted. Both systems are trained on LibriSpeech 100-hour data and the test-clean set is used for filtering.

This issue is illustrated by Fig. 1 which is a plot used for data selection for semi-supervised learning [20]. The plot shows that if all utterances are selected, then a conventional system and an attention-based sequence-to-sequence model have similar performance (7~8% WER). As the confidence threshold increases, the WER of the conventional system monotonically decreases. However, for the end-to-end model, a higher threshold does not always mean a reduced WER. The spike in Fig. 1 indicates that the end-to-end model is overconfident based on softmax probabilities.

To address the overconfidence issue, the *confidence estimation module* (CEM) is proposed for attention-based sequence-to-sequence models in Sec. 2. Sections 3 and 4 describes the setup and experiments that demonstrate the effectiveness of CEM. Section 5 shows the generalisation performance and impact on downstream tasks of the proposed method. Conclusions are drawn in Sec. 6.

\*Work was done while the author interned at Google.

## 2. CONFIDENCE ESTIMATION MODULE

An attention-based sequence-to-sequence model, such as the Listen, Attend and Spell (LAS) model [18], generally consists of an encoder, an attention mechanism and a decoder. As shown in the green block in Fig. 2, for an input utterance  $\mathbf{x}_1, \dots, \mathbf{x}_L$ , the encoder first transforms the input feature for each frame

$$\mathbf{e}_{1:L} = \text{ENCODER}(\mathbf{x}_{1:L}). \quad (1)$$

Then at a decoding step  $t$ ,

$$\mathbf{a}_t = \text{ATTENTION}(\mathbf{a}_{t-1}, \mathbf{d}_{t-1}, \mathbf{e}_{1:L}), \quad (2)$$

$$\mathbf{d}_t = \text{DECODER}(\mathbf{a}_t, \mathbf{d}_{t-1}, \text{EMB}(y_{t-1})), \quad (3)$$

$$p(y_t | y_{1:t-1}, \mathbf{x}_{1:L}) = \text{SOFTMAX}(\mathbf{d}_t). \quad (4)$$

where  $y$  is a label token normally representing a grapheme or a word piece [15] for ASR tasks. At the token level, confidence scores for ASR are defined as the probability of the token being correct. If the recogniser is very confident about the output token, then the corresponding confidence score should be close to 1. Word or utterance level confidence scores can be obtained by taking the average of tokens within a word or an utterance. For attention-based sequence-to-sequence models, each step in the auto-regressive decoder is treated as a classification task over all possible output tokens. However, there is a subtle difference between calibration for standard classification and confidence scores for sequences. For a hypothesis sequence, each token can either be correct, a substitution or an insertion. Because of the auto-regressive nature of the decoder and the use of teacher forcing approach for training, calibration behaviour for sequences with an incorrect history is uncertain. Furthermore, the model can be poorly calibrated when the model becomes very deep and large in pursuit of state-of-the-art performance [22].

To obtain high-quality confidence scores while keeping the model performance, the *confidence estimation module* (CEM) is proposed as shown in the red block in Fig. 2. The CEM is designed to be a lightweight module that can be easily configured on top of any attention-based sequence-to-sequence model. The CEM gathers information from the attention mechanism, the decoder state and the current token embedding which is fed into a fully connected (FC) layer. Then the sigmoid output layer generates a value between 0 and 1 that indicates the confidence score  $p_t$  for the current token, i.e.

$$p_t = \text{SIGMOID}(\text{FC}(\mathbf{a}_t, \mathbf{d}_t, \text{EMB}(y_t))). \quad (5)$$

Assuming there is an existing well-trained LAS model,  $n$ -best hypotheses can be generated by running a forward pass through the encoder and running a beam search through the decoder. Then the edit distance between each hypothesis sequence can be computed with respect to the ground truth reference sequence. The alignment from the edit distance computation can be used as the target for confidence if correct tokens are assigned as 1 while substituted or inserted tokens are assigned as 0. For example, if the ground truth sequence is “A B C D.” and one of the hypotheses is “A C C D.”, then the binary target sequence is  $\mathbf{c} = [1, 0, 1, 1]$ . For each of the  $n$ -best hypotheses, the CEM is trained to minimise the binary cross entropy between the estimated confidence  $\mathbf{p}$  and the target  $\mathbf{c}$ .

$$\mathcal{L}(\mathbf{c}, \mathbf{p}) = -\frac{1}{T} \sum_{t=1}^T \left( c_t \log(p_t) + (1 - c_t) \log(1 - p_t) \right). \quad (6)$$

The total loss for an utterance is the aggregated confidence estimation loss for all the  $n$ -best hypotheses. During training, all parameters for the attention-based sequence-to-sequence model are fixed.

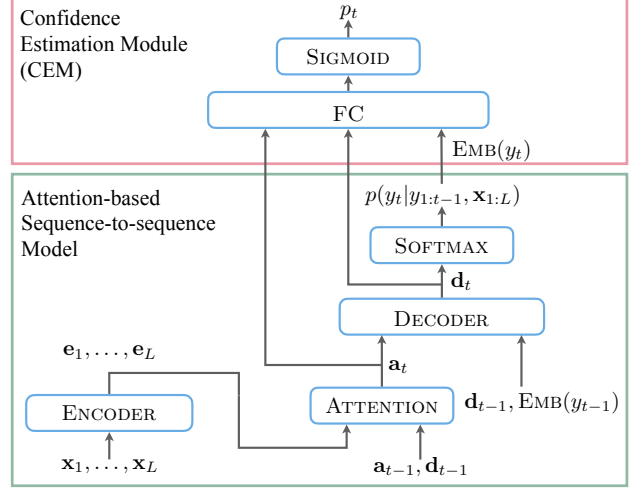


Fig. 2: Confidence estimation module (CEM) for attention-based sequence-to-sequence models.

## 3. EXPERIMENTAL SETUP

### 3.1. Evaluation Metrics

For each hypothesis, an alignment between the hypothesis and the corresponding ground truth can be obtained by computing the Levenshtein distance. If the hypothesis token is the same as the aligned reference token, the target confidence should be 1 and 0 otherwise.

To measure the quality of confidence scores, *normalised cross-entropy* (NCE) is a frequently used metric [23]. If we gather confidence scores for all tokens  $\mathbf{p} = [p_1, \dots, p_N]$  where  $p_n \in [0, 1]$  and their corresponding target confidence  $\mathbf{c} = [c_1, \dots, c_N]$  where  $c_n \in \{0, 1\}$ , NCE is given by

$$\text{NCE}(\mathbf{c}, \mathbf{p}) = \frac{H(\mathbf{c}) - H(\mathbf{c}, \mathbf{p})}{H(\mathbf{c})} \quad (7)$$

where  $H(\mathbf{c})$  is the entropy of the target sequence and  $H(\mathbf{c}, \mathbf{p})$  is the binary cross-entropy between the target and the estimated confidence scores. When confidence estimation is systematically better than the word correct ratio ( $\sum_{n=1}^N c_n / N$ ), NCE is positive. For perfect confidence scores, NCE is 1. NCE measures how close the confidence score is to the probability of the recognised word being correct.

However, in applications such as keyword spotting, it is the order of tokens ranked by confidence scores that matters. In these cases, operating points have to be chosen where hypotheses with confidence scores above a certain threshold  $\tilde{p}$  are deemed to be correct and incorrect otherwise. Precision-recall (P-R) curves are commonly used to illustrate the operating characteristics [24].

$$\text{precision}(\tilde{p}) = \frac{\text{TP}(\tilde{p})}{\text{TP}(\tilde{p}) + \text{FP}(\tilde{p})}, \quad \text{recall}(\tilde{p}) = \frac{\text{TP}(\tilde{p})}{\text{TP}(\tilde{p}) + \text{FN}(\tilde{p})}. \quad (8)$$

where TP is true positives, FP is false positives and FN is false negatives. Normally, when the threshold  $\tilde{p}$  increases, there are fewer false positives and more false negatives, which leads to higher precision and lower recall. The trade-off behaviour between precision and recall yields a downward trending curve from the top left corner to the bottom right corner. Therefore, the *area under the curve* (AUC) can measure the quality of the confidence estimator, which has a maximum value of 1. It is worth noting that two confidence estimators can have the same AUC value but different NCE values.

### 3.2. Data

The LibriSpeech train-clean-100 subset [25] is used for model training. The 100-hour subset reflects a typical use case of confidence scores where the WERs are moderately high and the amount of supervised data is limited. The dev and test sets are dev-clean/dev-other and test-clean/test-other sets. The input features are 80-dimension filterbank coefficients with  $\Delta$  and  $\Delta\Delta$ . The output targets are 16k word-pieces tokenised from the full LibriSpeech training set using a WPM model [26]. All experimental results in Sec. 4 on LibriSpeech are on the test-clean/test-other sets.

### 3.3. Baseline Model

The baseline model is an LAS model trained using the open-source Lingvo toolkit [27]. The encoder consists of a 2-layer convolutional neural network with max-pooling and stride of 2 and 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 in the baseline model is 184 million. The optimiser is Adam with learning rate 0.001 and the batch size is 512. During training, five regularisation techniques have been adopted, including dropout [28] of 0.1 on the decoder, uniform label smoothing [29] of 0.2, Gaussian weight noise [30] with zero mean and a standard deviation of 0.05 for all model parameters after 20k updates, SpecAugment [31] with 2 frequency masks with mask parameter  $F = 27$ , 2 time masks with mask parameter  $T = 40$ , and time warping with warp parameter  $W = 40$ , and an exponential moving average (EMA) [32] of all model parameters were used during training.

## 4. EXPERIMENTAL RESULTS

### 4.1. Effect of Regularisation Methods on Confidence Scores

Many training techniques have been developed for deep neural networks. They normally reduce the extent to which the model overfits to the training data and improve generalisation of the model. The state-of-the-art performance is often achieved when a large model is trained using aggressive regularisation [15]. As for the baseline setup described in Sec. 3.3, five techniques are used. Broadly speaking, regularisation methods can be classified into three categories, *i.e.* augmenting input features (SpecAugment), manipulating model weights (dropout, EMA & weight noise), and modifying output targets (label smoothing).

	WER ↓	AUC ↑	NCE ↑
baseline	7.5/21.6	0.976/0.912	-0.195/ 0.131
– dropout	7.8/22.0	0.977/0.916	-0.204/ 0.130
– EMA	8.2/24.8	0.974/0.903	-0.189/ 0.120
– label smoothing	10.6/24.6	0.985/0.950	0.106/-0.131
– weight noise	12.9/25.8	0.978/0.925	-0.459/-0.012
– SpecAugment	10.8/34.3	0.952/0.911	0.012/ 0.160

**Table 1:** ASR and token-level confidence performance by removing a regularisation method from the baseline model on test-clean/test-other. Confidence scores are based on softmax probabilities.

In Table 1, the ASR results and confidence performance of five additional models are shown, where each model is trained by removing one regularisation technique from the baseline setup. As expected, removing any of these regularisation methods results in an increased WER. However, the confidence performance may not

worsen when a regularisation method is excluded. Also, AUC and NCE values do not always change in the same direction. Although disentangling the contribution of each technique is challenging, Table 1 shows that augmenting input features and manipulating model weights can generally improve the confidence performance for at least one metric. Interestingly, by removing label smoothing, AUC is significantly better than the baseline and the gap of NCE values between test-clean and test-other sets is reversed. This shows that although softmax probabilities can be directly used as confidence scores, they can be heavily affected by regularisation techniques, especially if output targets are modified.

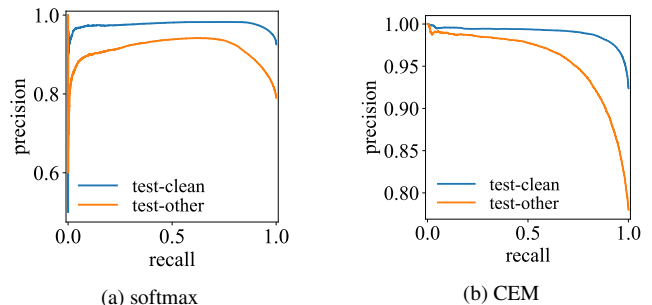
### 4.2. Confidence Estimation Module

To keep good ASR performance while having reliable confidence scores, a dedicated CEM can be trained as in Sec. 2. The CEM is added on top of the baseline model. During the CEM training, more aggressive SpecAugment (10 time masks with mask parameter  $T = 50$ ) is used to increase the WER on the training set for more negative training samples. For each utterance, 8-best hypotheses are generated on-the-fly and are aligned with the reference to obtain the binary training targets. The CEM only has one fully-connected layer with 256 units. The number of additional parameters is 0.4% of the baseline LAS model. Piece-wise linear mappings (PWLMs) [33] are estimated on dev-clean/dev-other and are then applied to test-clean/test-other, so that the confidence scores better match with token or word accuracy. Since PWLM is monotonic, NCE is boosted while AUC remains unchanged as the relative order of confidence scores is unchanged.

		AUC ↑	NCE ↑ (w/o PWLM)	NCE ↑ (w/ PWLM)
token	softmax	0.976/0.912	-0.195/0.131	0.166/0.172
	CEM	<b>0.990/0.958</b>	0.189/0.019	<b>0.344/0.275</b>
word	softmax	0.981/0.927	-0.180/0.139	0.269/0.195
	CEM	<b>0.990/0.962</b>	0.192/0.039	<b>0.350/0.270</b>

**Table 2:** Comparison of confidence scores between using softmax probabilities and using the CEM on the baseline model. The first row corresponds to the baseline in Table 1.

Table 2 reports the confidence metrics at the token level and the word level. The word-level confidence is the average of the token-level ones if a word consists of multiple tokens. The AUC is improved at both token and word levels by using CEM. Unlike softmax probabilities, NCE values are all positive for CEM. After PWLM, CEM yields much higher NCE values than softmax probabilities.



**Fig. 3:** Precision-recall curves for token-level confidence scores on LibriSpeech test-clean and test-other sets.

However, AUC values do not show the whole picture. As shown in Fig. 3, the P-R curves of softmax and CEM are drastically different. A sharp downward spike at the high-confidence region in Fig. 3(a) corresponds to a low precision and a low recall. In other words, softmax probabilities are overconfident for some incorrect tokens, which also explains the spike shown in Fig. 1. CEM, however, suffers little from overconfidence as Fig. 3(b) depicts the desired trade-off between precision and recall. Overall, the CEM is a more reliable confidence estimator under both metrics.

#### 4.3. Effect of Language Model Fusion on Confidence Scores

For end-to-end ASR, shallow fusion of a language model (LM) [34] is commonly used to improve ASR performance during decoding. The effect of shallow fusion on confidence estimation is investigated. The language model used is a three-layer LSTM network with width 4096 trained on the LibriSpeech LM corpus, which shares the same word-piece vocabulary as the attention-based sequence-to-sequence model. To take LM information for confidence estimation, the input to the CEM is extended by the LM probability for the current token. Other setups are the same as Sec. 4.2.

		WER ↓	AUC ↑	NCE ↑
baseline	softmax	7.5/21.6	0.981/0.927	0.269/0.195
	CEM		<b>0.990/0.962</b>	<b>0.350/0.270</b>
+ LM	softmax	6.8/19.8	0.981/0.928	0.103/0.109
	CEM		<b>0.991/0.966</b>	<b>0.337/0.263</b>

**Table 3:** ASR and word-level confidence performance for models with and without RNNLM shallow fusion (with PWLM).

Table 3 shows the word-level confidence scores after PWLM for both softmax probabilities and CEM. Although WERs on test-clean and test-other decreased by 8~9% relatively, there is no clear improvement on AUC and even substantial degradation for NCE. By comparing the first and second blocks of Table 3, the CEM improves the quality of confidence estimation more significantly when an additional LM is used. The contrast of P-R curves between softmax and CEM with LM shallow fusion is similar to Fig. 3.

## 5. ANALYSIS

#### 5.1. Generalisation to a Mismatched Domain

Since CEM is a model-based approach and the training data for CEM is the same as the ASR, the model is naturally more confident on the training set. Although the mismatch between training and test for CEM is mitigated by having more aggressive augmentation during training and applying PWLMs estimated on dev sets during testing, it is still unclear how well the confidence scores from CEM generalises to data from a mismatched domain.

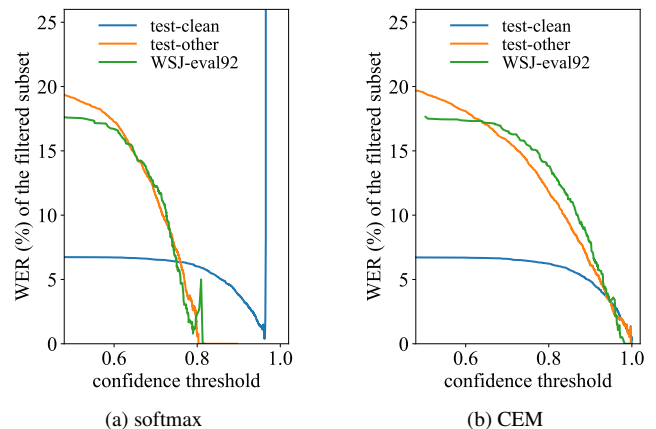
		WER ↓	AUC ↑	NCE ↑
baseline	softmax	18.7	0.935	0.230
	CEM		<b>0.970</b>	<b>0.280</b>
+ LM	softmax	17.7	0.933	0.159
	CEM		<b>0.965</b>	<b>0.266</b>

**Table 4:** ASR and confidence performance on WSJ eval92. PWLM is estimated on LibriSpeech dev-other set. LM used is trained on LibriSpeech LM corpus as in Sec. 4.3.

Wall Street Journal (WSJ) [35], a dataset of clean read speech of news articles, is in a moderately mismatched domain compared to LibriSpeech in terms of speaker, style and vocabulary. In Table 4, WSJ eval92 test set is fed into the same setup as in Sec. 4.3, where all models are trained on LibriSpeech. Similar to observations in Table 3, shallow fusion worsens the confidence estimation by softmax probabilities despite reduced WER. CEM improves the quality of confidence estimation significantly with or without LM.

#### 5.2. Implications for Downstream Tasks

Confidence scores are widely used to select unlabelled data for semi-supervised learning [20] to improve ASR performance. First, a speech recogniser is trained using the limited transcribed data. Then the recogniser transcribes the unlabelled data, which can be taken as noisy labels to train the existing model further. However, erroneous automatic transcription can hurt the model. Assuming confidence scores can reflect the word error rate well, filtering out utterances with low confidence can be beneficial to semi-supervised training.



**Fig. 4:** WERs of filtered utterances w.r.t. confidence thresholds for softmax and CEM with LM shallow fusion.

Similar to Fig. 1 and plots used for semi-supervised learning [20], Fig. 4 shows the WER of the filtered utterances whose confidence scores are above the corresponding threshold. If confidence scores strongly correlate with WER, a higher threshold will filter a subset with lower WER. In Fig. 4(a), sharp spikes at the high confidence region clearly indicates overconfidence based on softmax probabilities. In contrast, curves for all three test sets in Fig. 4(b) are monotonically going down without spikes, which shows that confidence scores by CEM match WER much more closely.

## 6. CONCLUSIONS

Using softmax probabilities of attention-based sequence-to-sequence models for confidence estimation is not reliable. To this end, the lightweight confidence estimation module (CEM) is proposed. The effectiveness of CEM is demonstrated with and without shallow fusion of a language model. Furthermore, the performance of CEM can generalise well on a slightly mismatched domain and may benefit various downstream tasks, such as adaptation [7], semi-supervised training [20] and model combination [36, 37]. A similar module to CEM can also be introduced for deletion prediction as deletion errors are not within the scope of confidence estimation but can be important for many applications [12, 38]. With minimal modification, CEM is also applicable to recurrent neural network transducer (RNN-T) for speech recognition [39].

## 7. REFERENCES

- [1] F. Wessel, R. Schlüter, K. Macherey, & H. Ney, “Confidence measures for large vocabulary continuous speech recognition,” *IEEE Trans. on Speech and Audio Processing*, vol. 9, pp. 288–298, 2001.
- [2] H. Jiang, “Confidence measures for speech recognition: A survey,” *Speech Communication*, vol. 45, pp. 455–470, 2005.
- [3] D. Yu, J. Li, & L. Deng, “Calibration of confidence measures in speech recognition,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, pp. 2461–2473, 2011.
- [4] R.H.Y. Chan & P.C. Woodland, “Improving broadcast news transcription by lightly supervised discriminative training,” *Proc. ICASSP*, Montreal, 2004.
- [5] G. Tür, D.Z. Hakkani-Tür, & R. Schapire, “Combining active and semi-supervised learning for spoken language understanding,” *Speech Communication*, vol. 45, pp. 171–186, 2005.
- [6] G. Riccardi & D. Hakkani-Tür, “Active learning: Theory and applications to automatic speech recognition,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, pp. 504–511, 2005.
- [7] L.F. Uebel & P.C. Woodland, “Speaker adaptation using lattice-based MLLR,” *Proc. ITRW on Adaptation Methods for Speech Recognition*, Sophia Antipolis, 2001.
- [8] G. Evermann & P.C. Woodland, “Posterior probability decoding, confidence estimation and system combination,” *Proc. NIST Speech Transcription Workshop*, College Park, 2000.
- [9] L. Mangu, E. Brill, & A. Stolcke, “Finding consensus in speech recognition: Word error minimization and other applications of confusion networks,” *Computer Speech and Language*, vol. 14, pp. 373–400, 2000.
- [10] M.S. Seigel & P.C. Woodland, “Combining information sources for confidence estimation with CRF models,” *Proc. Interspeech*, Florence, 2011.
- [11] K. Kalgaonkar, C. Liu, Y. Gong, & K. Yao, “Estimating confidence scores on ASR results using recurrent neural networks,” *Proc. ICASSP*, Brisbane, 2015.
- [12] A. Ragni, Q. Li, M.J.F. Gales, & Y. Wang, “Confidence estimation and deletion prediction using bidirectional recurrent neural networks,” *Proc. SLT*, Athens, 2018.
- [13] Q. Li, P. Ness, A. Ragni, & M.J.F. Gales, “Bi-directional lattice recurrent neural networks for confidence estimation,” *Proc. ICASSP*, Brighton, 2019.
- [14] A. Kastanos, A. Ragni, & M.J.F. Gales, “Confidence estimation for black box automatic speech recognition systems using lattice recurrent neural networks,” *Proc. ICASSP*, Barcelona, 2020.
- [15] C.C. Chiu, T. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, *et al.*, “State-of-the-art speech recognition with sequence-to-sequence models,” *Proc. ICASSP*, Calgary, 2018.
- [16] Y. He, T. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, *et al.*, “Streaming end-to-end speech recognition for mobile devices,” *Proc. ICASSP*, Brighton, 2019.
- [17] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, & Y. Bengio, “Attention-based models for speech recognition,” *Proc. NIPS*, Montreal, 2015.
- [18] W. Chan, N. Jaitly, Q.V. Le, & O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” *Proc. ICASSP*, Shanghai, 2016.
- [19] M. Zpotoczny, P. Pietrzak, A. Lancucki, & J. Chorowski, “Lattice generation in attention-based speech recognition models,” *Proc. Interspeech*, Graz, 2019.
- [20] D. Park, Y. Zhang, Y. Jia, W. Han, C.C. Chiu, B. Li, *et al.*, “Improved noisy student training for automatic speech recognition,” *Proc. Interspeech*, Shanghai, 2020.
- [21] D. Hendrycks & K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” *Proc. ICLR*, Toulon, 2017.
- [22] C. Guo, G. Pleiss, Y. Sun, & K. Weinberger, “On calibration of modern neural networks,” *Proc. ICML*, Sydney, 2017.
- [23] M. Siu, H. Gish, & F. Richardson, “Improved estimation, evaluation and applications of confidence measures for speech recognition,” *Proc. Eurospeech*, Rhodes, 1997.
- [24] J. Davis & M. Goadrich, “The relationship between precision-recall and ROC curves,” *Proc. ICML*, Pittsburgh, 2006.
- [25] V. Panayotov, G. Chen, D. Povey, & S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” *Proc. ICASSP*, Brisbane, 2015.
- [26] M. Schuster & K. Nakajima, “Japanese and Korean voice search,” *Proc. ICASSP*, Kyoto, 2012.
- [27] J. Shen, P. Nguyen, Y. Wu, Z. Chen, M.X. Chen, Y. Jia, *et al.*, “Lingvo: A modular and scalable framework for sequence-to-sequence modeling,” *ArXiv.org*, 1902.08295, 2019.
- [28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, & R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, & Z. Wojna, “Rethinking the inception architecture for computer vision,” *Proc. CVPR*, Las Vegas, 2016.
- [30] A. Graves, “Practical variational inference for neural networks,” *Proc. NIPS*, Granada, 2011.
- [31] D. Park, W. Chan, Y. Zhang, C.C. Chiu, B. Zoph, E.D. Cubuk, *et al.*, “SpecAugment: A simple data augmentation method for automatic speech recognition,” *Proc. Interspeech*, Graz, 2019.
- [32] B. Polyak & A. Juditsky, “Acceleration of stochastic approximation by averaging,” *Siam Journal on Control and Optimization*, vol. 30, pp. 838–855, 1992.
- [33] G. Evermann & P.C. Woodland, “Large vocabulary decoding and confidence estimation using word posterior probabilities,” *Proc. ICASSP*, Istanbul, 2000.
- [34] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.C. Lin, *et al.*, “On using monolingual corpora in neural machine translation,” *ArXiv.org*, 1503.03535, 2015.
- [35] D.B. Paul & J. Baker, “The design for the Wall Street Journal-based CSR corpus,” *Proc. ICSLP*, Banff, 1992.
- [36] Q. Li, C. Zhang, & P.C. Woodland, “Integrating source-channel and attention-based sequence-to-sequence models for speech recognition,” *Proc. ASRU*, Singapore, 2019.
- [37] J. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” *Proc. ASRU*, Santa Barbara, 1997.
- [38] M.S. Seigel & P.C. Woodland, “Detecting deletions in ASR output,” *Proc. ICASSP*, Florence, 2014.
- [39] A. Graves, “Sequence transduction with recurrent neural networks,” *Proc. ICML Workshop on Representation Learning*, Edinburgh, 2012.