# Multi-Task Learning for End-to-End ASR Word and Utterance Confidence with Deletion Prediction

**David Qiu  Yanzhang He  Qiujia Li[1]  Yu Zhang  Liangliang Cao  Ian McGraw**

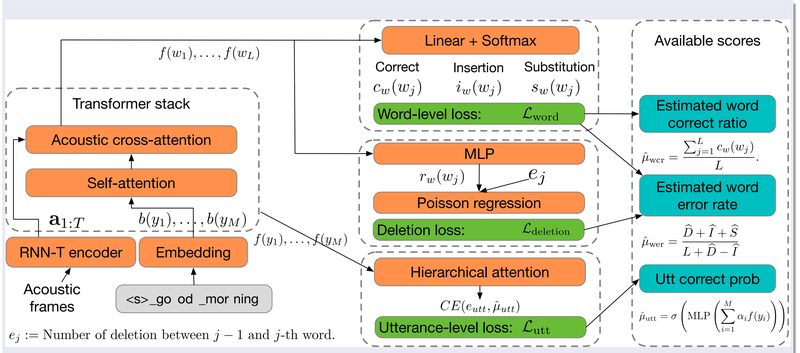
Google LLC, USA  [1]University of Cambridge, UK

{qdavid,yanzhanghe}@google.com

## 1. Summary

- Many prior works on ASR confidence focused on token or word confidence
- Models correct, insertion, and substitution, but ignores deletions
- We incorporate deletion signals through explicitly predicting deletion lengths or binary classification of sentence correctness
- Multi-task training improves word and utterance confidence metrics
- Improved confidence scores can be used for rescoring to lower WER

## 2. Model Architecture




- Transformer based per-wordpiece feature extractor
- Utterance-level feature extractor based on hierarchical attention [3]
- Three prediction branches trained with different losses:

**Word-Level Confidence Estimation** [2]:
Predict the whether each hypothesized word is correct, insertion, or substitution

$$\mathcal{L}_{\text{word}} = -\sum_{j=1}^{L}\left[\mathbb{1}_j^{\text{cor}} \log c_w(w_j) + \mathbb{1}_j^{\text{ins}} \log i_w(w_j) + \mathbb{1}_j^{\text{sub}} \log s_w(w_j)\right]$$

$$\hat{\mu}_{\text{wcr}} = \frac{\sum_{j=1}^{L} c_w(w_j)}{L}$$

**Deletion Length Prediction**:
Predict the deletion length between each pair of hypothesized words

$$e_j = \begin{cases} \#(\text{del}) \text{ before the first word}, & \text{if } j = 1 \\ \#(\text{del}) \text{ after the last word}, & \text{if } j = L+1 \\ \#(\text{del}) \text{ between } j-1 \text{ and } j\text{-th word}, & \text{otherwise.} \end{cases}$$

$$\mathcal{L}_{\text{deletion}} = -\sum_{j=1}^{L+1}\left[e_j r_w(w_j) - \exp\left(r_w(w_j)\right)\right]$$

$$\hat{\mu}_{\text{wer}} = \frac{\widehat{D} + \widehat{I} + \widehat{S}}{L + \widehat{D} - \widehat{I}}$$

**Utterance-Level Confidence Estimation**:
Predict whether the utterance is recognized with zero WER

$$e_{\text{utt}} = \begin{cases} 1, & \text{if utterance WER} = 0 \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathcal{L}_{\text{utt}} = -[e_{\text{utt}} \log \hat{\mu}_{\text{utt}} + (1 - e_{\text{utt}}) \log(1 - \hat{\mu}_{\text{utt}})]$$

## 3. Multi-Task Learning

- Generate *N*-best hypotheses using frozen ASR recognizing audio with SpecAugment and multi-condition training (MTR) [1]
- Run Levenshtein edit distance between the ground truth and each *N*-best to obtain labels

$$\mathcal{L}_{\text{total}} = \frac{1}{L}\mathcal{L}_{\text{word}} + \frac{\lambda_{\text{deletion}}}{L+1}\mathcal{L}_{\text{deletion}} + \lambda_{\text{utt}}\mathcal{L}_{\text{utt}}$$

Table: Training objectives for five proposed models.

| Model | Model Size | Loss | | |
|---|---|---|---|---|
| | | $\mathcal{L}_{\text{word}}$ | $\mathcal{L}_{\text{deletion}}$ | $\mathcal{L}_{\text{utt}}$ |
| W | 16.6M | ✓ | ✗ | ✗ |
| U | 17.0M | ✗ | ✗ | ✓ |
| WD | 16.9M | ✓ | ✓ | ✗ |
| WU | 17.2M | ✓ | ✗ | ✓ |
| WUD | 17.5M | ✓ | ✓ | ✓ |

## 4. Experimental Results

- Depending on which loss is applied, there are multiple confidence scores available
- $c_w$ is always used for word confidence
- $\hat{\mu}_{\text{utt}}$ estimates the utterance probability of correct, and should be used for AUC (ranking)
- $\hat{\mu}_{\text{wcr}}$ and $\hat{\mu}_{\text{wer}}$ estimates the WER, and should be used for RMSE eval when available

Table: Utterance confidence scores' priorities by metrics.

| Model | Available Scores | AUC-ROC & PR | RMSE |
|---|---|---|---|
| W | $\hat{\mu}_{\text{wcr}}$ | $\hat{\mu}_{\text{wcr}}$ | $\hat{\mu}_{\text{wcr}}$ |
| U | $\hat{\mu}_{\text{utt}}$ | $\hat{\mu}_{\text{utt}}$ | N/A |
| WD | $\hat{\mu}_{\text{wcr}}$, $\hat{\mu}_{\text{wer}}$ | $1 - \hat{\mu}_{\text{wer}}$ | $1 - \hat{\mu}_{\text{wer}}$ |
| WU | $\hat{\mu}_{\text{wcr}}$, $\hat{\mu}_{\text{utt}}$ | $\hat{\mu}_{\text{utt}}$ | $\hat{\mu}_{\text{wcr}}$ |
| WUD | $\hat{\mu}_{\text{wcr}}$, $\hat{\mu}_{\text{utt}}$, $\hat{\mu}_{\text{wer}}$ | $\hat{\mu}_{\text{utt}}$ | $1 - \hat{\mu}_{\text{wer}}$ |

Voice Search

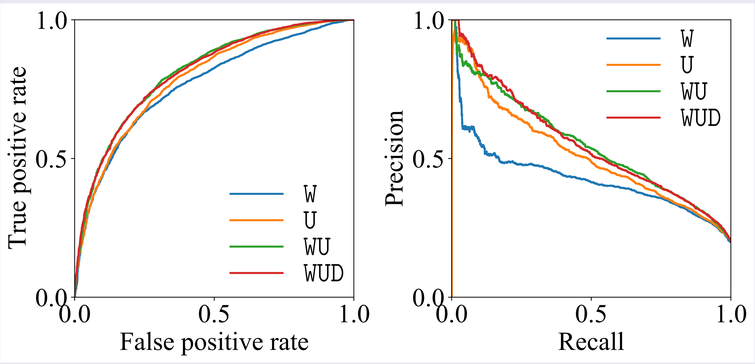
| Model | Word-level NCE | Word-level AUC ROC | Word-level AUC PR | Utterance-level AUC ROC | Utterance-level AUC PR | Utterance-level (1-WER) RMSE |
|---|---|---|---|---|---|---|
| W | 0.348 | 0.927 | 0.451 | 0.765 | 0.428 | 0.226 |
| U | – | – | – | 0.788 | 0.515 | – |
| WD | 0.350 | 0.928 | 0.459 | 0.767 | 0.440 | 0.216 |
| WU | **0.378** | **0.931** | **0.489** | **0.810** | 0.549 | 0.223 |
| WUD | 0.365 | 0.929 | 0.476 | **0.810** | **0.552** | **0.213** |

Long-tail Maps

| Model | Word-level NCE | Word-level AUC ROC | Word-level AUC PR | Utterance-level AUC ROC | Utterance-level AUC PR | Utterance-level (1-WER) RMSE |
|---|---|---|---|---|---|---|
| W | 0.285 | 0.875 | 0.494 | 0.704 | 0.549 | 0.301 |
| U | – | – | – | 0.721 | 0.593 | – |
| WD | 0.291 | 0.875 | 0.504 | 0.704 | 0.558 | 0.296 |
| WU | **0.304** | **0.883** | **0.524** | **0.755** | **0.636** | 0.296 |
| WUD | **0.304** | 0.882 | 0.516 | 0.746 | 0.634 | **0.292** |

## 5. Experimental Results (cont.)

Figure: Utterance-level ROC and PR curves on Voice Search for Models W, U, WU, WUD.




(a) ROC cuves.  (b) PR curves.

Table: Rescoring WERs (%) showing the WER improvement of confidence reranking over RNN-T top hypothesis.

| Model | Rank | Voice Search | Long-tail Maps |
|---|---|---|---|
| RNN-T | Beam search | 6.4 | 14.0 |
| W | $\hat{\mu}_{\text{wcr}}$ | 6.2 | 13.8 |
| U | $\hat{\mu}_{\text{utt}}$ | 6.3 | 14.4 |
| WU | $\hat{\mu}_{\text{wcr}}$ | **6.1** | **13.6** |
| WU | $\hat{\mu}_{\text{utt}}$ | **6.1** | 14.1 |
| WUD | $\hat{\mu}_{\text{wcr}}$ | 6.2 | **13.6** |
| WUD | $\hat{\mu}_{\text{utt}}$ | 6.2 | 14.1 |

## 6. Conclusions

- Deletion signals are valuable for helping word confidence models
- Performing multi-task training requires minimal additional capacity and improves word and utterance confidence metrics
- Confidence models can serve as a light-weight alternative to a full size rescorer

## 7. References

[1] C. Kim, A. Misra, K. K. Chin, T. Hughes, A. Narayanan, et al. Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home. In *Interspeech*, 2017.

[2] D. Qiu, Q. Li, Y. He, Y. Zhang, B. Li, L. Cao, R. Prabhavalkar, D. Bhatia, W. Li, K. Hu, et al. Learning word-level confidence for subword end-to-end ASR. In *ICASSP*, 2021.

[3] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *NAACL*, 2016.