KU LEUVEN

Yichen Jia
Hugo Van hamme

Departement Electrical Engineering ESAT-PSI
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

Paper Poster Code

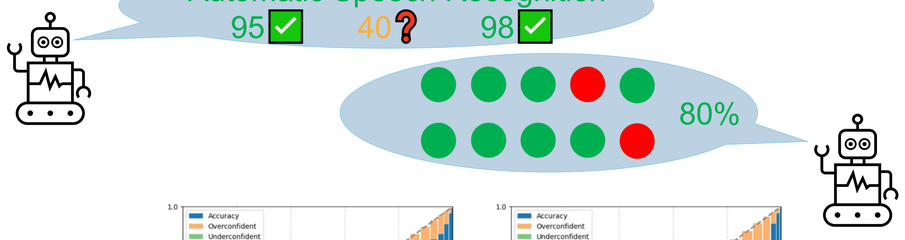
Leveraging Beam Search Information for Confidence Estimation in E2E ASR

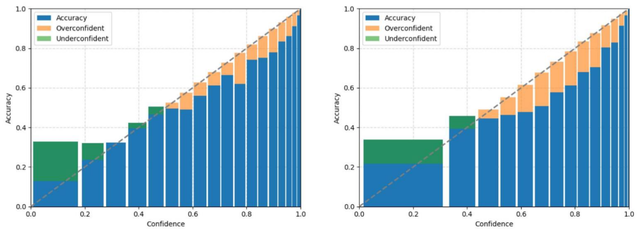
Summary

To estimate **confidence** for **E2E ASR** systems, recent research has proposed Confidence Estimation Modules that incorporate features from the backbone ASR model, but most existing approaches remain **architecture-dependent**.

We propose **Score-Rank Confidence Estimation Module (SR-CEM)**, a module that leverages **beam search information** to generate token- and word-level confidence scores. Experiments show that SR-CEM achieves effective calibration on in-domain and out-of-domain English data, and demonstrates robustness across hybrid and transducer ASR architectures, multiple decoding strategies, and Dutch, noisy, and conversational speech conditions. SR-CEM consistently reduces **Maximum Calibration Error** across diverse evaluation conditions while remaining **architecture-independent**.

Problem Definition



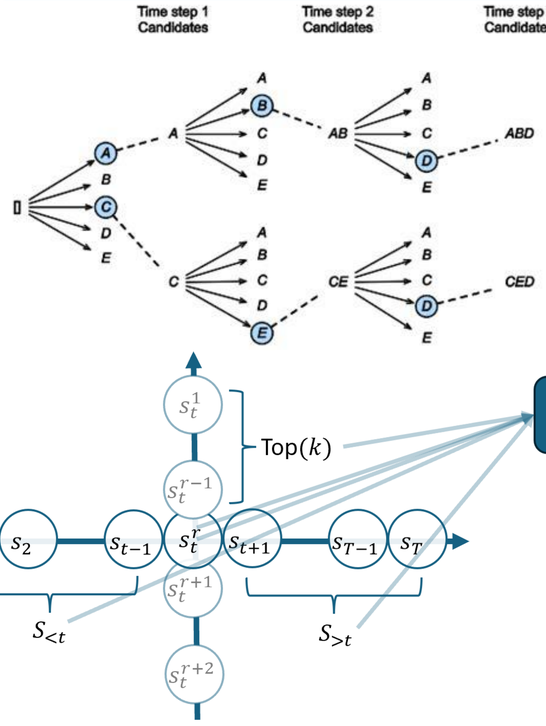


Token Level Word Level

Aim

1. Reliable confidence scores for E2E ASR at **token** and **word** level;
2. Without altering the ASR **accuracy**;
3. Only use the information available in **beam search**.

Proposed: SR-CEM



$$\hat{c}_Y(y_t) = \sigma\left(W_{Y,2} \cdot \mathrm{ReLU}(W_{Y,1}\mathbf{f}_Y(y_t) + \mathbf{b}_{Y,1}) + \mathbf{b}_{Y,2}\right)$$

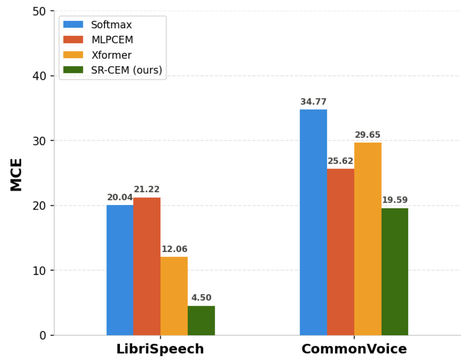
$$\mathbf{f}_Y(y_t) = \left[s_Y(y_t),\ r(y_t),\ S_{<t},\ S_{>t},\ \mathrm{Topk}(t)\right]$$

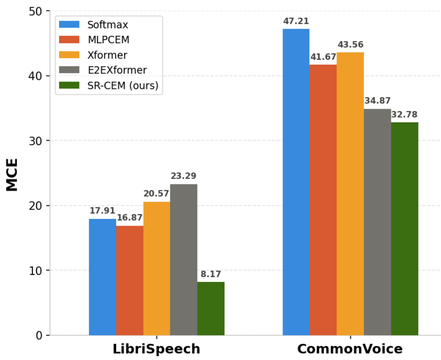
$$\hat{c}_W(w_l) = \sigma\left(W_{W,2} \cdot \mathrm{ReLU}(W_{W,1}\mathbf{f}_W(w_l) + \mathbf{b}_{W,1}) + \mathbf{b}_{W,2}\right)$$

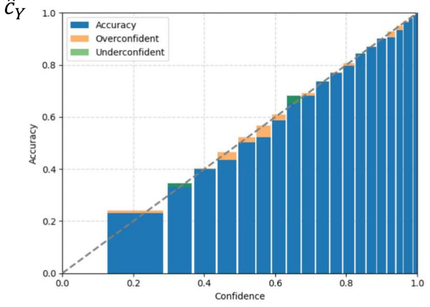
$$\mathbf{f}_W(w_l) = [s_W(w_l),\ r_l(w_l),\ S_{<l},\ S_{>l},\ Q_l]$$

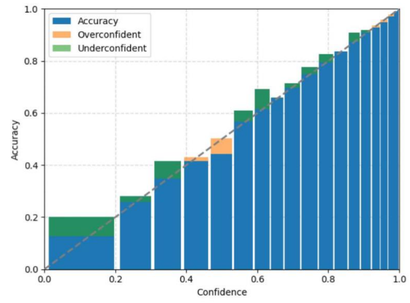
1. SR-CEM is a **two-layer perceptron**;
2. Collect the **rank and score** information in beam search;
3. Then form token or word level features;
4. Train a lightweight Confidence Estimation Module (CEM) on these features;
5. Supervised training is done using **cross entropy loss** between correctness and confidence score.

Results: Token and Word level









Token Level Word Level

Conclusion

SR-CEM achieves **best MCE** among all baselines

Ablation confirms each component contributes

Generalizes across architectures, decoding strategies & diverse speech conditions

Reference

1. Q. Li et al., “Confidence estimation for attention-based sequence-to-sequence models for speech recognition,” in Proc. ICASSP IEEE Int. Conf. Acoust., Speech Signal Process., 2021, pp. 6388–6392.
2. D. Qiu et al., “Learning word-level confidence for subword end-to-end ASR,” in Proc. ICASSP IEEE Int. Conf. Acoust., Speech Signal Process., 2021, pp. 6393–6397.
3. N. Ravi et al., “TeLeS: Temporal lexeme similarity score to estimate confidence in end-to-end ASR,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 32, pp. 4399–4408, 2024.