ANALYZING UNCERTAINTIES IN SPEECH RECOGNITION USING DROPOUT

Apoorv Vyas, Pranay Dighe, Sibo Tong, Hervé Bourlard
IDIAP Research Institute, Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
(apoorv.vyas, sibo.tong, bourlard)@idiap.ch, pdighe@apple.com

OVERVIEW

Goals
• Confidence Estimation
• Error Localization
• Automatic performance measurement without transcripts

Prior-Art
• Lattice-based approaches.
• Classifier-based approaches.

Our approach
• Use dropout for bayesian uncertainty approximation in DNN output.

ESTIMATING UNCERTAINTY IN ASR OUTPUT

Figure 1: Decoding with dropout on at test time. Each network represents a different random selection of the active nodes. The white nodes denote dropped out units.

Confidence = \[ \sum_{i=1}^{N} I(D_{\text{act}}^{[w]} = D_{\text{act}}^{[w]}) / N \]

WER = \[ E_{w} / T_{w} \]

RESULT: N-BEST VS DROPOUT

Figure 2: Comparing N-best list against the dropout samples for test utterance: oh i see uh-huh

• N-best list contains several outputs even when the best path is correct.
• Dropout samples contain the same output (high confidence).

RESULTS (ERROR LOCALIZATION)

Table 1: Error Localization comparison using IoU metric

• Overall IoU of ~ 0.6 means significant number of predicted errors are accurate.
• IoU metric is much higher for shorter utterances.

RESULTS (WER ESTIMATION)

Table 2: Results on estimating WER using dropout uncertainty.

• N-best estimation overestimates WER on short utterances and underestimates WER on longer utterances.
• N-best list contains all different hypothesis thus cannot estimate WER as 0.
• All dropout hypotheses can be identical when model is confident.

CONCLUSIONS & FUTURE WORK

• Variations in different hypotheses with dropout are often highly localized at certain word positions and depict locations of potential errors.
• Experiments with CTC and DNN-HMM acoustic models show that our approach accurately estimates word error rates and word confidences and is more robust to the utterance length, compared to lattice-based approaches.
• In future, we intend to use word-level predictive uncertainty in the output for model combination and for semi-supervised learning.

ACKNOWLEDGMENT

This research was supported by Swiss National Science Foundation project SHBSM, grant agreement 200021-175388, and the European Community H2020 SUMMA project No. 688139.