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

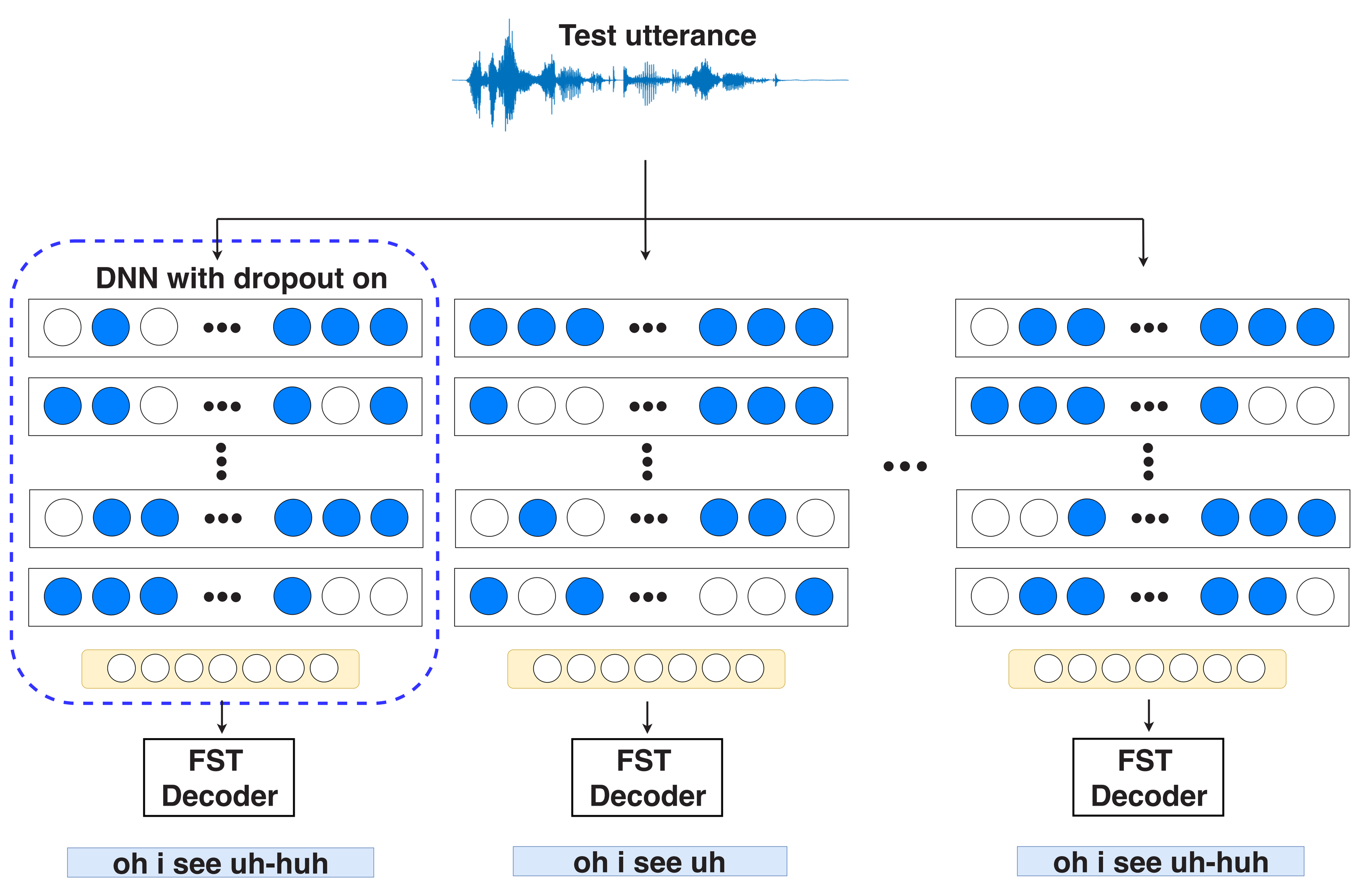

Figure 1: Decoding with dropout on at test time. Each network represents a different random selection of the active nodes. The white nodes denote dropped out units.

$$\textbf{Confidence} = \frac{\sum_{k=1}^{N} I(D_{\text{on}}^{k}[w] = D_{\text{off}}[w])}{N} \qquad \textbf{WER} = \frac{E_{\mu}}{L_{\mu}}$$

Figure.2 Example of predicting ASR errors using dropout uncertainty.

D_{off} :	i	i	agree with	the	a hundred percent there	or
GT :	i	say	agree with	you	a hundred percent there	—

****

D_{on}^{1} :	i	i	agree with	you	a hundred percent there	—
D_{on}^{2} :	yes	i	agree with	the	a hundred percent there	—
D_{on}^{3} :	i	i	agree with	you	a hundred percent there	or
D_{on}^{4} :	yes	i	agree with	the	a hundred percent there	—
C_w :	0.5	1	1 1	0.5	1 1 1 1	0.25

True Positive False Positive Missed Detection

Experiment Set up

Dataset: Switchboard (110h subset)

Acoustic Models:

- DNN-HMM: Kaldi nnet1, 6 layers, 2048 neurons, 0.2 dropout
- CTC: 4 BLSTM layers, 320 cells, 0.2 dropout

Metric:

- Intersection-over-Union (IoU) for error localization
- Relative difference for WER estimation

Result: N-best vs Dropout

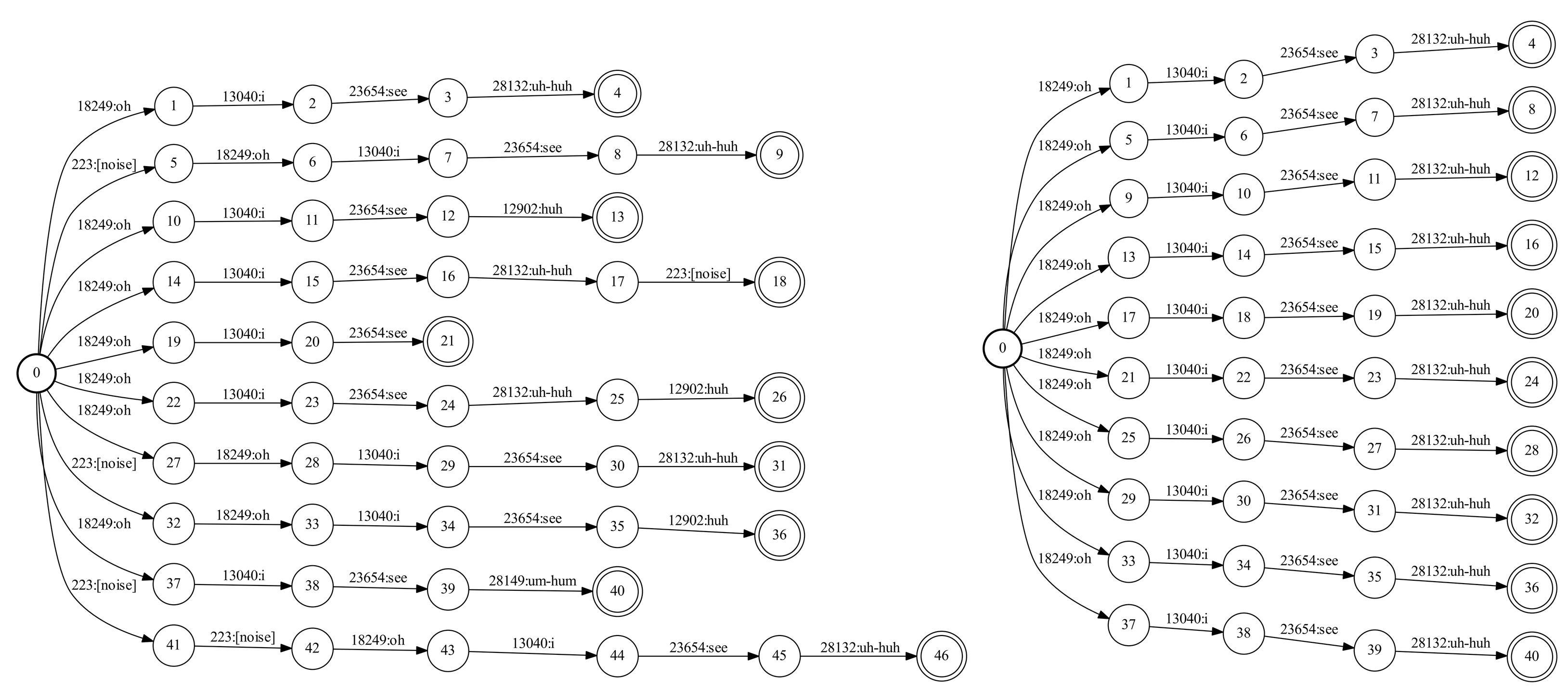

Figure 2: Comparing N-best list against the dropout samples for test utterance: *oh i see uh-huh*

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

Results (Error Localization)

ASR System	S.L [1-3]	S.L [4-6]	S.L [7-10]	S.L [11-Max]	S.L [1-Max]
DNN-*dev*-Dr	0.78	**0.68**	0.55	**0.42**	**0.55**
DNN-*dev*-Nb	0.62	0.49	0.44	0.38	0.45
DNN-*dev*-Ka	**0.79**	0.67	**0.56**	**0.42**	0.54
DNN-*test*-Dr	**0.83**	**0.59**	**0.55**	**0.42**	**0.59**
DNN-*test*-Nb	0.66	0.51	0.45	0.39	0.50
DNN-*test*-Ka	0.82	0.56	0.51	0.41	0.58
CTC-*dev*-Dr	**0.79**	**0.62**	**0.58**	**0.45**	**0.56**
CTC-*dev*-Nb	0.51	0.46	0.42	0.36	0.41
CTC-*dev*-Ka	0.78	0.57	0.50	0.39	0.51
CTC-*test*-Dr	**0.84**	**0.62**	**0.53**	**0.44**	**0.61**
CTC-*test*-Nb	0.45	0.50	0.45	0.35	0.41
CTC-*test*-Ka	0.83	0.56	0.50	0.40	0.58

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)

ASR System	S.L [1-3]	S.L [4-6]	S.L [7-10]	S.L [11-Max]	S.L [1-Max]
DNN-*test*-Gt	35.5	28.8	25.1	22.3	23.3
DNN-*test*-Dr	**33.2**	**31.0**	**25.2**	**21.5**	**22.7**
DNN-*test*-Nb	73.6	42.1	30.2	13.7	19.7
CTC-*test*-Gt	30.6	31.2	25.3	23.3	24.0
CTC-*test*-Dr	**34.4**	**35.5**	**29.7**	**23.9**	**25.2**
CTC-*test*-Nb	93.1	49.0	34.3	15.9	22.7

Table 2: Results on estimating WER using dropout uncertainty. *S.L.* refers to sentence length. *Dr*: dropout estimate, *Nb*: N-best list estimate, and *Gt*: ground truth word error rate.

- 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.

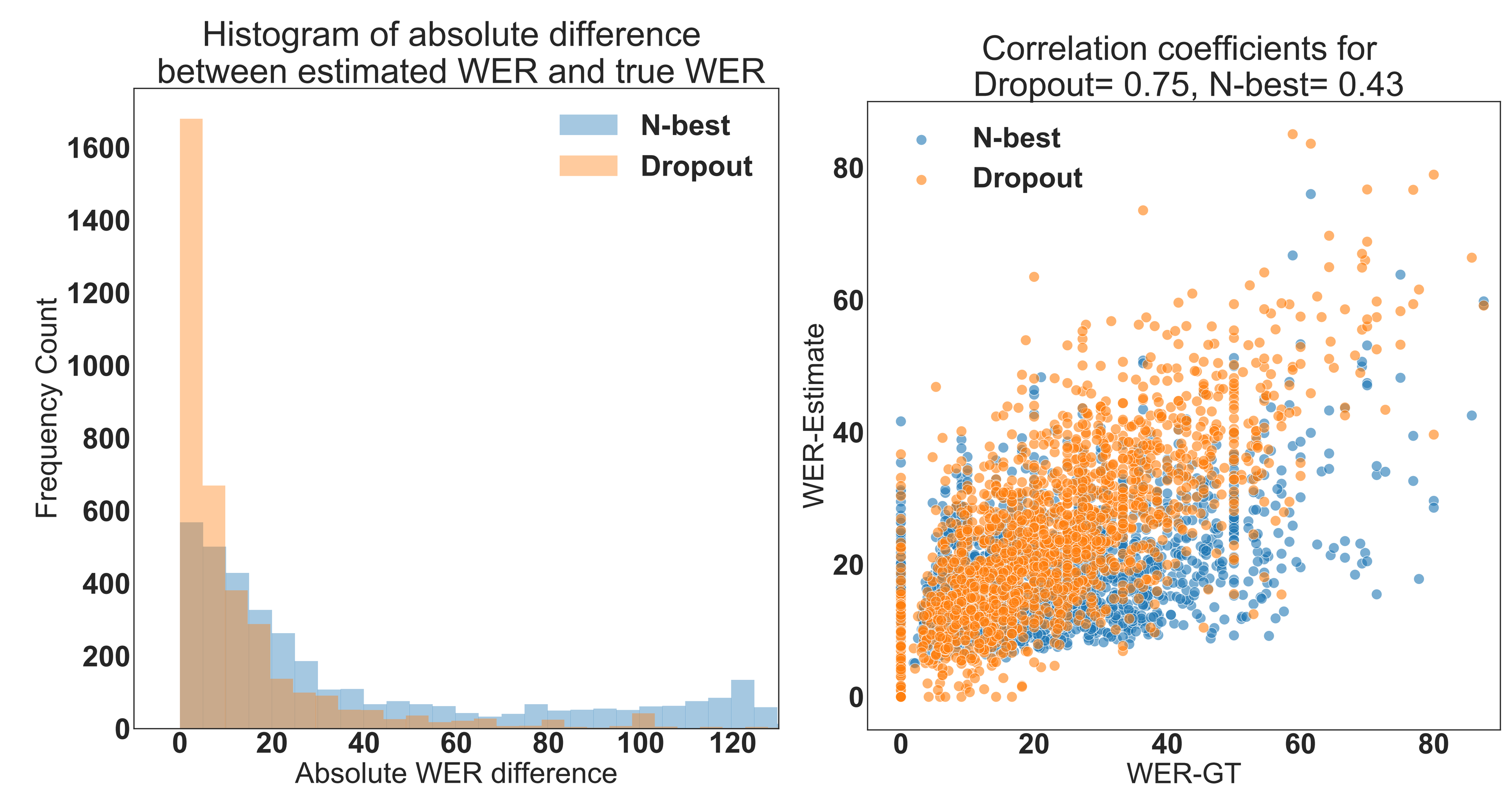

Figure 3: Comparison between dropout and N-best list based WER estimation for CTC system. (a) Histogram of absolute difference between estimated WER and true WER. (b) Correlation between estimated WER and true WER.

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 SHISSM, grant agreement 200021-175589, and the European Community H2020 SUMMA project No. 688139