Ensemble combination between different time segmentations

Jeremy H. M. Wong, Dimitrios Dimitriadis, Kenichi Kumatani, Yashesh Gaur, George Polovets, Partha Parthasarathy, Eric Sun, Jinyu Li, and Yifan Gong

Microsoft speech and language group

jeremy.wong@microsoft.com

MOTIVATION

Task: Speech recognition

Problem

Hypothesis-level combination requires all models to use the same input time segmentations.

Proposal

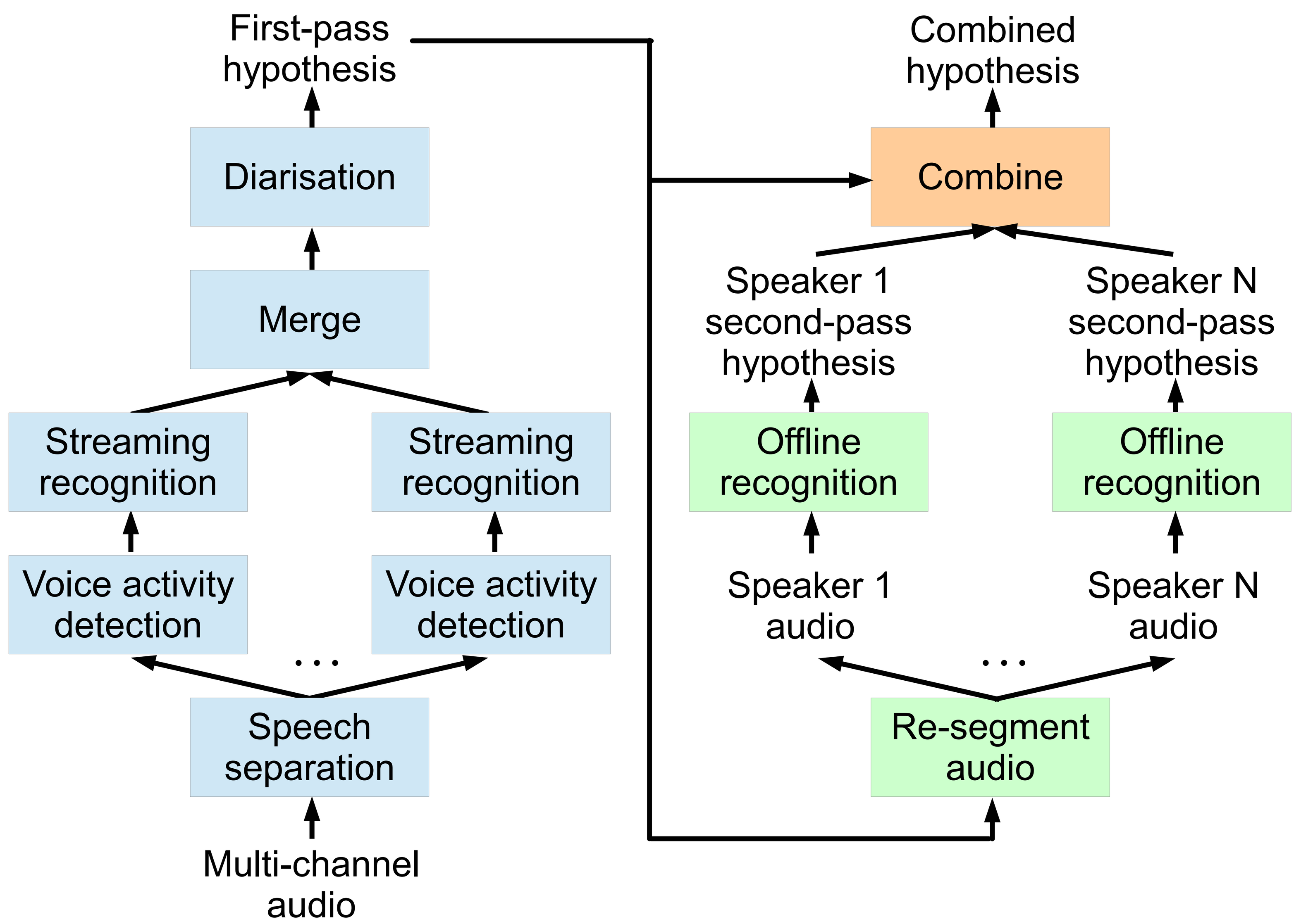
Allow different time segmentations between models by splitting and re-joining the hypothesis N-best lists.

Applications

Allow combinations between:

- Different voice activity detection front-ends.
- Different unsynchronised recording devices.
- Overlapping inference.
- 1st pass used to refine time segmentation of 2nd pass.

MEETING TRANSCRIPTION SETUP



1st pass streaming ASR → diarisation → 2nd pass offline ASR

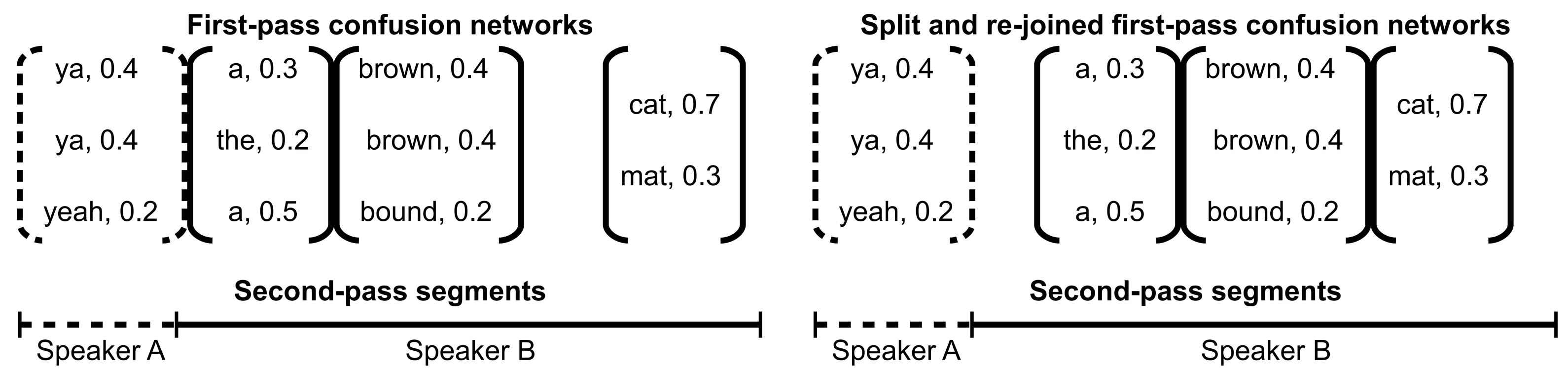
- 1st pass ASR uses VAD segments.
- 2nd pass ASR uses per-speaker segments from diarisation.
- Want to combine 1st pass and 2nd pass hypotheses.

Data:

- *dev* - 51 meetings, 23 hours
- *eval* - 60 meetings, 35 hours
- Average of 7 participants per meeting

MULTI-PASS COMBINATION

CONFUSION NETWORK SPLITTING



Steps:

1. Convert N-best list into confusion network.
2. Estimate start and end times of confusion sets.
3. Estimate confusion set speaker from 1-best hypothesis.
4. Split up confusion network into separate confusion sets.
5. Re-join consecutive confusion sets of the same speaker.
6. Do Confusion Network Combination (CNC).

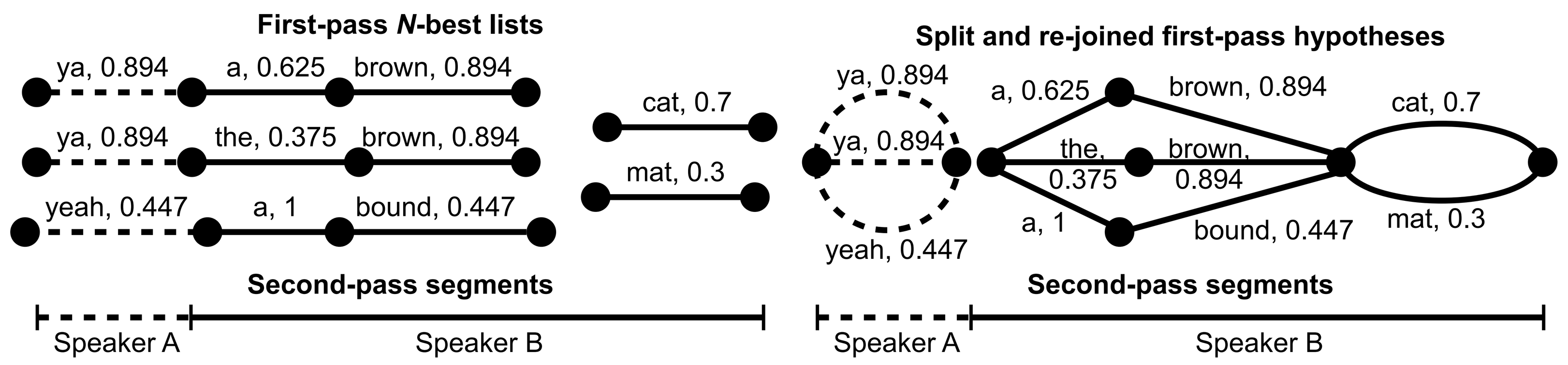
Advantages:

- 1-best is preserved after re-joining.

Disadvantages:

- Confusion set times are approximate.
- Context of language model scores is not preserved.

N-BEST LIST SPLITTING



Steps:

1. Distribute hypothesis scores to words.
2. Estimate speakers for N-best words from 1-best hypothesis.
3. Split up the N-best lists.
4. Re-join N-best lists according to segment time and speaker.
5. Do Minimum Bayes' Risk (MBR) combination.

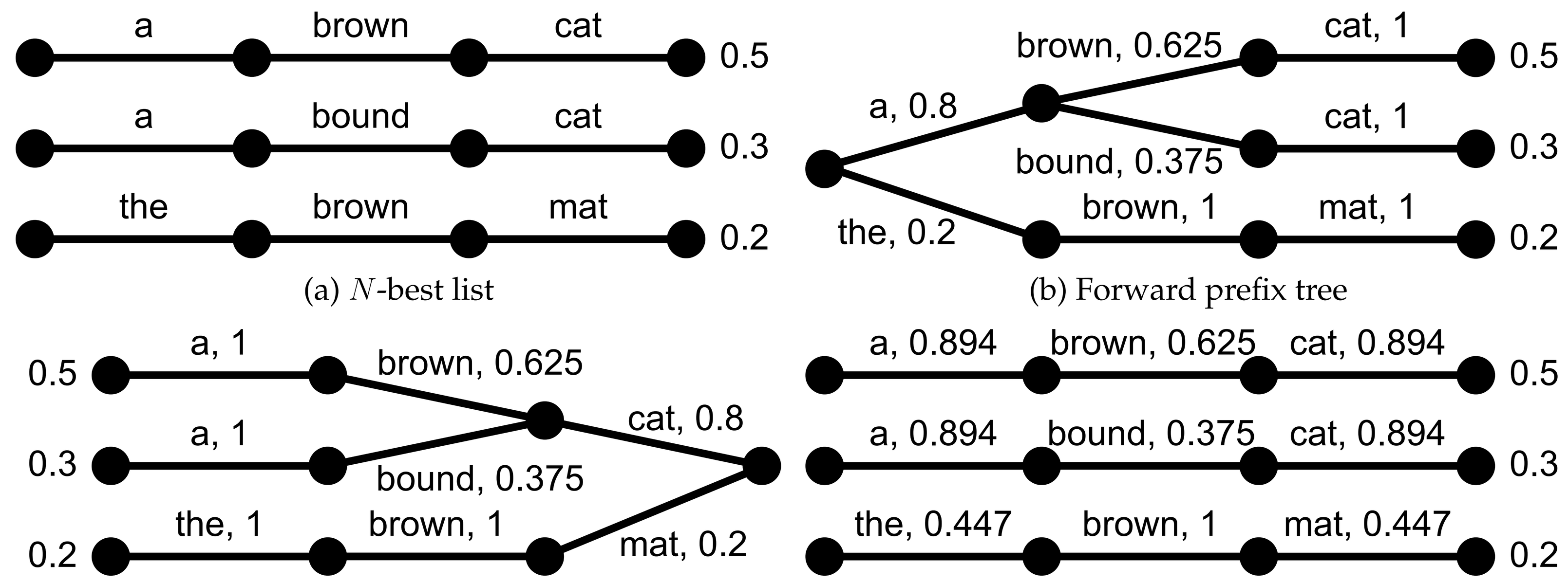
Advantages:

- Exact word start and end times are preserved.
- Context of language model scores is preserved.

Disadvantages:

- Hypothesis rank order may not be preserved after re-joining.

HYPOTHESIS SCORES TO WORD SCORES



(a) N-best list (b) Forward prefix tree (c) Backward suffix tree (d) Average scores from both trees

- Black-box speech recogniser may not produce per-word scores.
- Want to estimate per-word scores from per-hypothesis scores.

Steps:

1. Convert N-best list into prefix and suffix trees.
2. Push weights to branches.
3. Take log-average of scores from prefix and suffix trees.

EXPERIMENTS

Distribution of hypothesis scores to words, on 1st pass *eval*

Split	Per-word scores	Speaker-attributed WER (%)
no	original	20.43
yes	original	22.09
	language model re-score	22.09
	prefix tree	20.62
	suffix tree	20.60
	average	20.55

- Best performance with average of prefix and suffix trees.

Multi-pass combination (Speaker-attributed WER (%))

	dev	*eval*
1st pass streaming hybrid	21.43	20.43
2nd pass offline hybrid	19.93	19.13
2nd pass offline LAS	19.91	19.04
CNC streaming hybrid + offline hybrid	20.01	19.10
CNC streaming hybrid + offline LAS	19.71	18.71
MBR streaming hybrid + offline hybrid	19.83	19.00
MBR streaming hybrid + offline LAS	19.30	18.43
MBR offline hybrid + offline LAS	19.11	18.24

- N-best list splitting outperforms confusion network splitting.
- Combination with no increase in 2nd pass computational cost.
- Hybrid + LAS outperforms hybrid + hybrid.

CONCLUSION

- Distribute hypothesis scores to words using trees.
- Combine different time segments by splitting N-best lists.