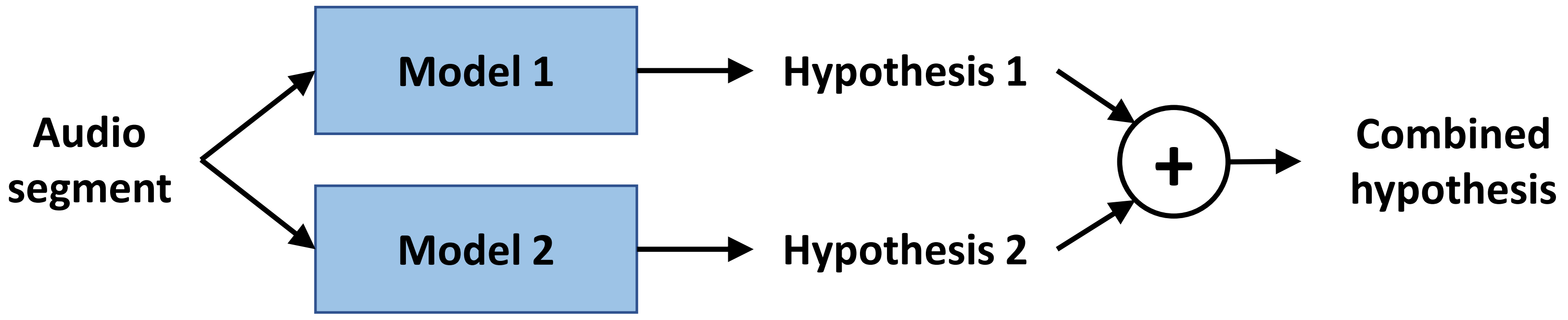
Microsoft

Ensemble combination between different time segmentations

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

Microsoft Speech and Language Group

Ensemble combination



- Hypothesis-level combination assumes that all models use the same input time segments.

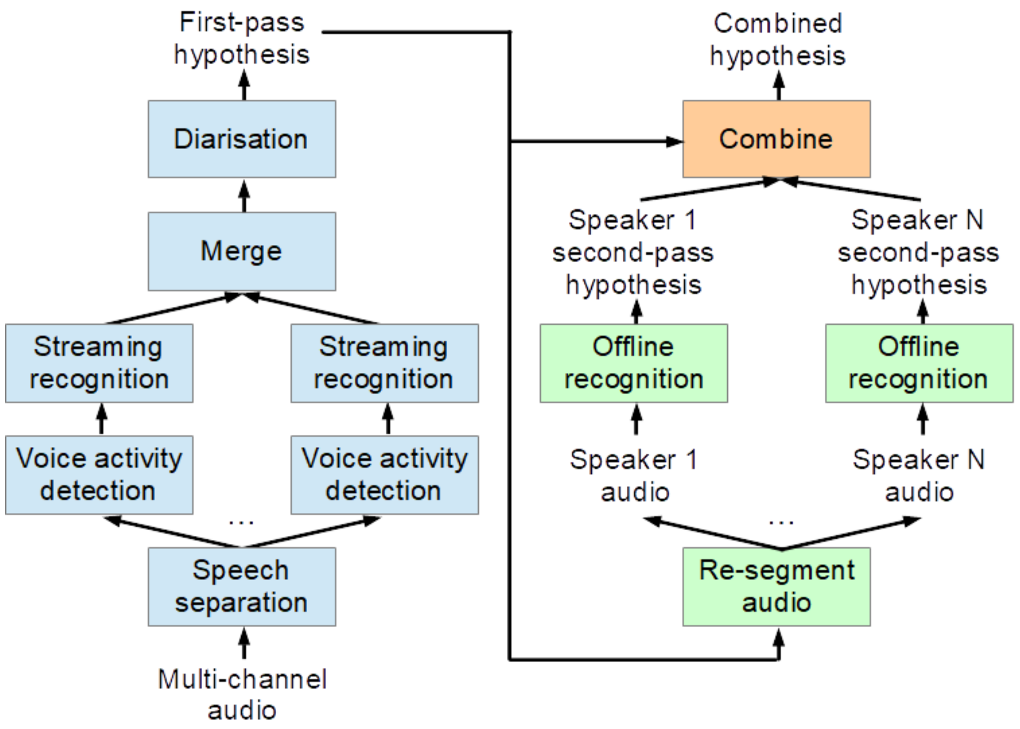
Propose:

- Method to allow different input segmentation times between models.

Applications for different time segmentations

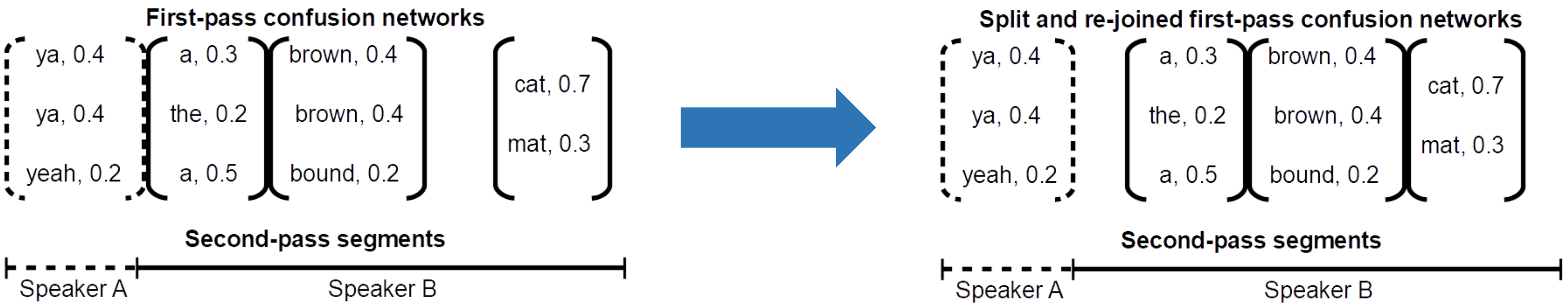
- Combination between different time segmentations can be used for:
  - Different VAD front-ends for each model.
  - Audio from multiple unsynchronised recording devices.
  - Overlapping inference.
  - Using a 1st pass ASR to refine the time segmentations for a 2nd pass ASR.

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.
- Want to combine 1st pass and 2nd pass ASR hypotheses to improve 2nd pass performance.

Confusion network splitting



1. Convert N-best list into confusion network.
2. Estimate start and end times of each confusion set.
3. Estimate the speaker ID for each confusion set from the 1-best hypothesis.
4. Split up confusion network into separate confusion sets.
5. Re-join consecutive confusion sets to match time segments.
6. Do Confusion Network Combination (CNC) between all models.

T. Yoshioka et. al., "*Meeting transcription using virtual microphone arrays*", Tech. Rep., Microsoft, May 2019

Confusion network splitting

Advantages:

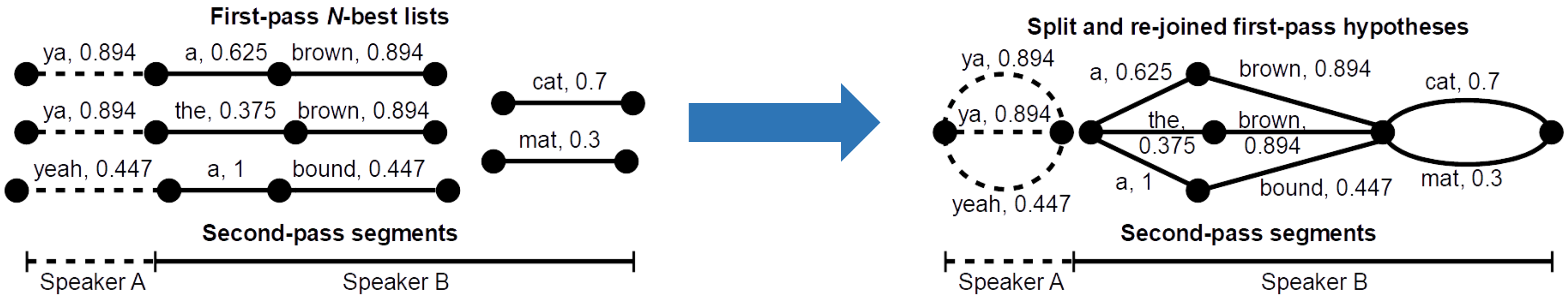
- 1-best is preserved after splitting and re-joining.

Disadvantages:

- Start and end times of each confusion set are approximate.
- Word sequence context of language model scores is not preserved.

Microsoft

N-best list splitting



1. Distribute hypothesis scores to words.
2. Estimate the speaker ID for each N-best word from the 1-best hypothesis.
3. Split up the N-best lists.
4. Re-join N-best lists according to segment times.
5. Do Minimum Bayes' Risk (MBR) combination between all models.

N-best list splitting

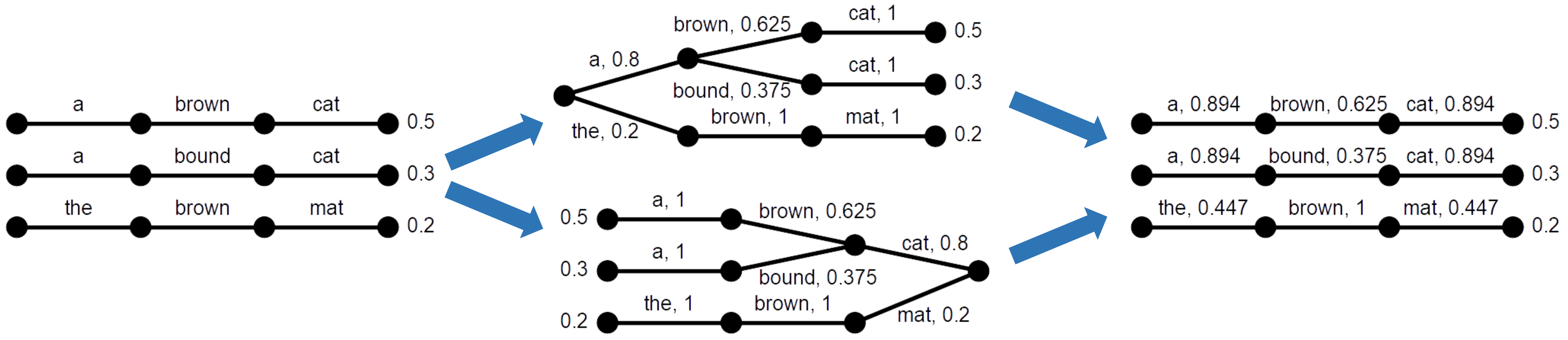
Advantages:

- Exact word start and end times are preserved from ASR decoding.
- Word sequence context of language model scores is preserved.

Disadvantages:

- 1-best may not be preserved after splitting and re-joining.

Distribute hypothesis scores to words



- Black-box ASR system may only produce per-hypothesis scores.
- Estimate per-word scores by:
  1. Convert N-best list to prefix and suffix trees.
  2. Push weights to branches.
  3. Take log-average of per-word scores from prefix and suffix trees.
- Prefix and suffix trees concentrate weights at opposite ends.

Experiments

Dataset:

- Internal Microsoft meetings.
- *dev* set: 51 meetings, 23 hours
- *eval* set: 60 meetings, 35 hours
- *Average of 7 participants per meeting.*

Speaker-attributed WER Metric:

- For each speaker, compute the WER of that speaker’s hypothesis vs reference.
- Average the WERs over all speakers.

Experiments

Models:

- **1st pass hybrid:** streaming latency-controlled and layer-trajectory BLSTM.
- **2nd pass hybrid:** ensemble of 2 offline BLSTMs.
- **2nd pass LAS:** offline BLSTM encoder, LSTM decoder.
- **Hybrid LM:** 5-gram + NNLM

N-best list size: 16

Microsoft

Score distribution method

- Distribute hypothesis-level scores to words for streaming 1st pass model.
- Split and re-join 1st pass N-best lists to match 2nd pass segments.

Split	Per-word scores	*eval* 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
	log-average	20.55

- After splitting, log-average between prefix and suffix trees performs best.
- Splitting yields degradation.

Microsoft

Multi-pass combination

- Single model performance.

		Speaker-attributed WER (%)	
Segments	**Model**	***dev***	***eval***
1st pass	streaming hybrid	21.43	20.43
2nd pass	streaming hybrid	20.87	19.96
	offline hybrid	19.93	19.13
	offline LAS	19.91	19.04

- Offline model outperforms streaming model.
- 2nd pass segments yield gains over 1st pass segments for the same model.

Microsoft

Multi-pass combination

Single model

Segments	Model	Speaker-attributed WER (%) *dev*	*eval*
1st pass	streaming hybrid	21.43	20.43
2nd pass	streaming hybrid	20.87	19.96
	offline hybrid	19.93	19.13
	offline LAS	19.91	19.04

Combination between 1st and 2nd pass hypotheses

Combination	Speaker-attributed WER (%) *dev*	*eval*
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

- MBR with N-best splitting outperforms CNC with confusion network splitting.
- Offline hybrid + offline LAS performs best, but is computationally expensive.
- Streaming hybrid + offline LAS yields reasonable gains, with only single model in 2nd pass.
- Hybrid + LAS outperforms hybrid + hybrid, suggesting greater diversity.
- Streaming hybrid (on 2nd pass segments) + offline hybrid *eval* WER = 18.37 %.

Summary

- **Proposed:**
  - Allow different time segments in combination by splitting and re-joining of N-best lists.
  - Estimate per-word scores from per-hypothesis scores using trees.

- Improve 2^{nd} pass performance without additional computational cost.

- Showed that hybrid + LAS outperforms hybrid + hybrid.