Imperial College London

EUSIPCO 2020

January 2021

Analysis of Phonetic Dependence of Segmentation Errors in Speaker Diarization

Simon W. McKnight, Aidan O. T. Hogg, Patrick A. Naylor

Speaker Diarization

Background

- Distinguishes speakers speaking at different times
  - usually unsupervised
  - often referred to as “who spoke when”
- Operates on speech signals:
  - feature extraction on frames
  - grouping features from multiple consecutive frames
  - clustering/labelling
  - resegmentation

Evaluation/Scoring

- Standard scoring methodology based on `md-eval.pl` v.22
  - unchanged for ~15 years

$$\underbrace{DER}_{\text{diarization error rate}} = \frac{\overbrace{FA}^{\text{false alarm}} + \overbrace{MISS}^{\text{missed speaker}} + \overbrace{ERROR}^{\text{wrong speaker}}}{\underbrace{TOTAL}_{\text{total speech time}}}$$

- Possible relaxations:
  - forgiveness collars around ground truth speaker boundaries
  - removal overlapping speaker segments

Imperial College London

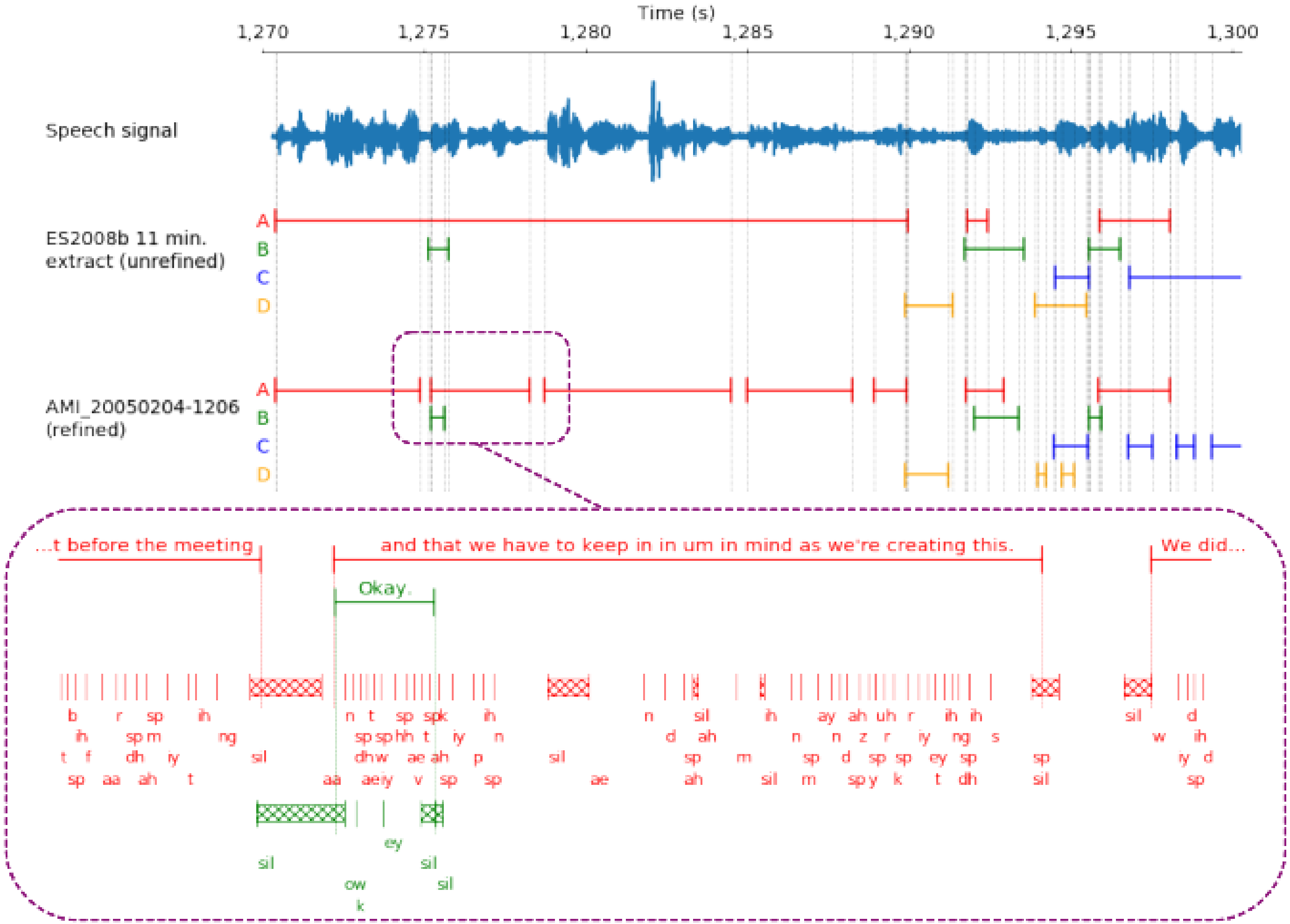
Research Aims

- Recent speaker diarization challenges (e.g. DIHARD I, II and III) have removed forgiveness collars
  - speaker diarization systems should rightly be evaluated on their entire performance
  - … but inherent uncertainty and subjectivity in ground truth speaker segmentation could unfairly penalize systems that correctly estimate speaker segment boundaries
- – DiarTk and `md-eval.pl` with original v refined ground truth segments AMI_20050204-1206 and 11 mins of ES2008b gives DERs:

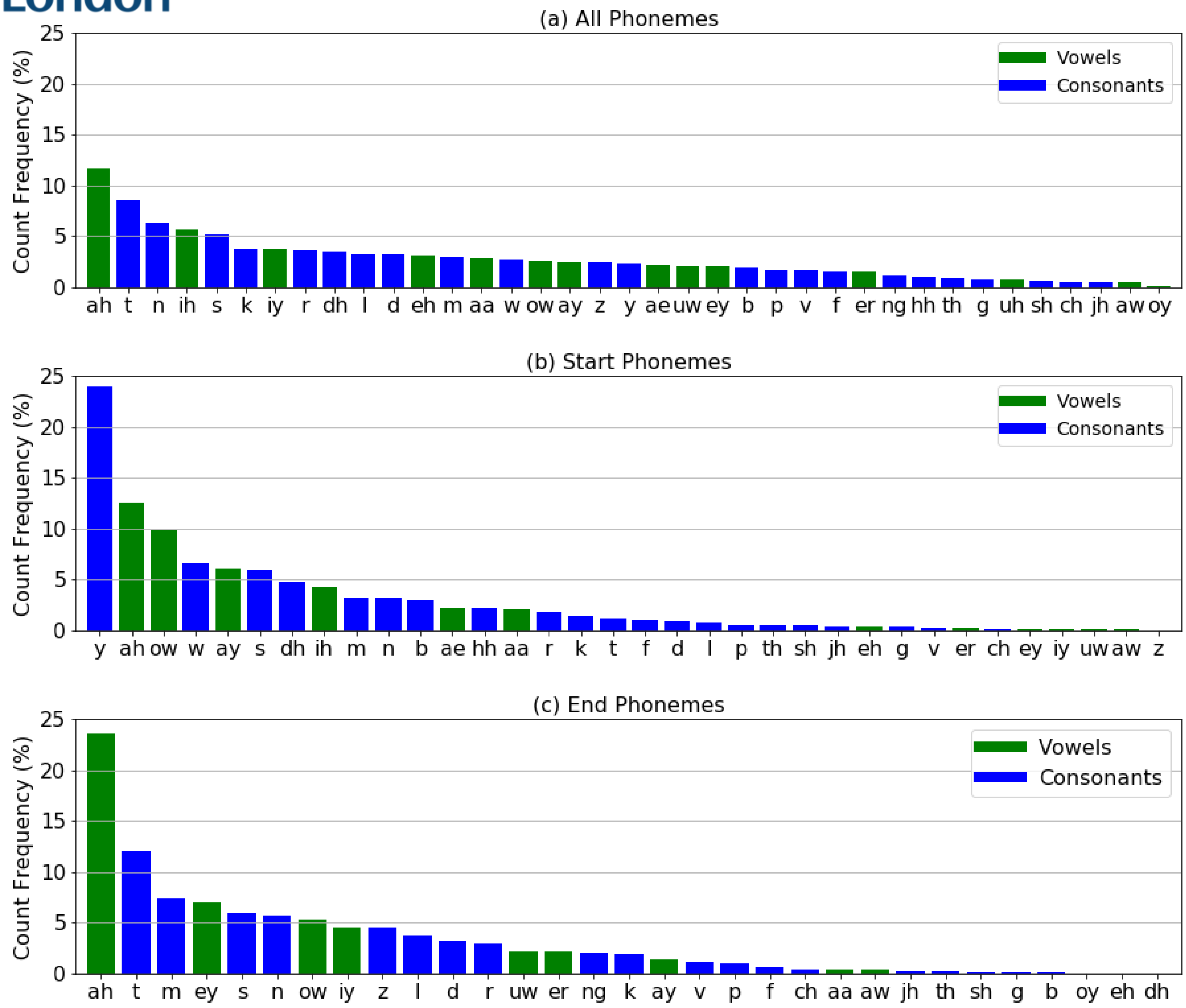
Collar (+/- ms)	Original → Original	Refined → Refined	Original → Refined
250	11.51%	8.79%	19.39%
0	21.26%	20.23%	29.14%

- This research investigates phoneme dependence of uncertainty in AMI Corpus
  - note these are original versions of ground truth segments
  - quantification of uncertainty and its effect on scoring

Example Differences in Ground Truth Segments



AMI Corpus Phonemes


(a) All Phonemes

(b) Start Phonemes

(c) End Phonemes

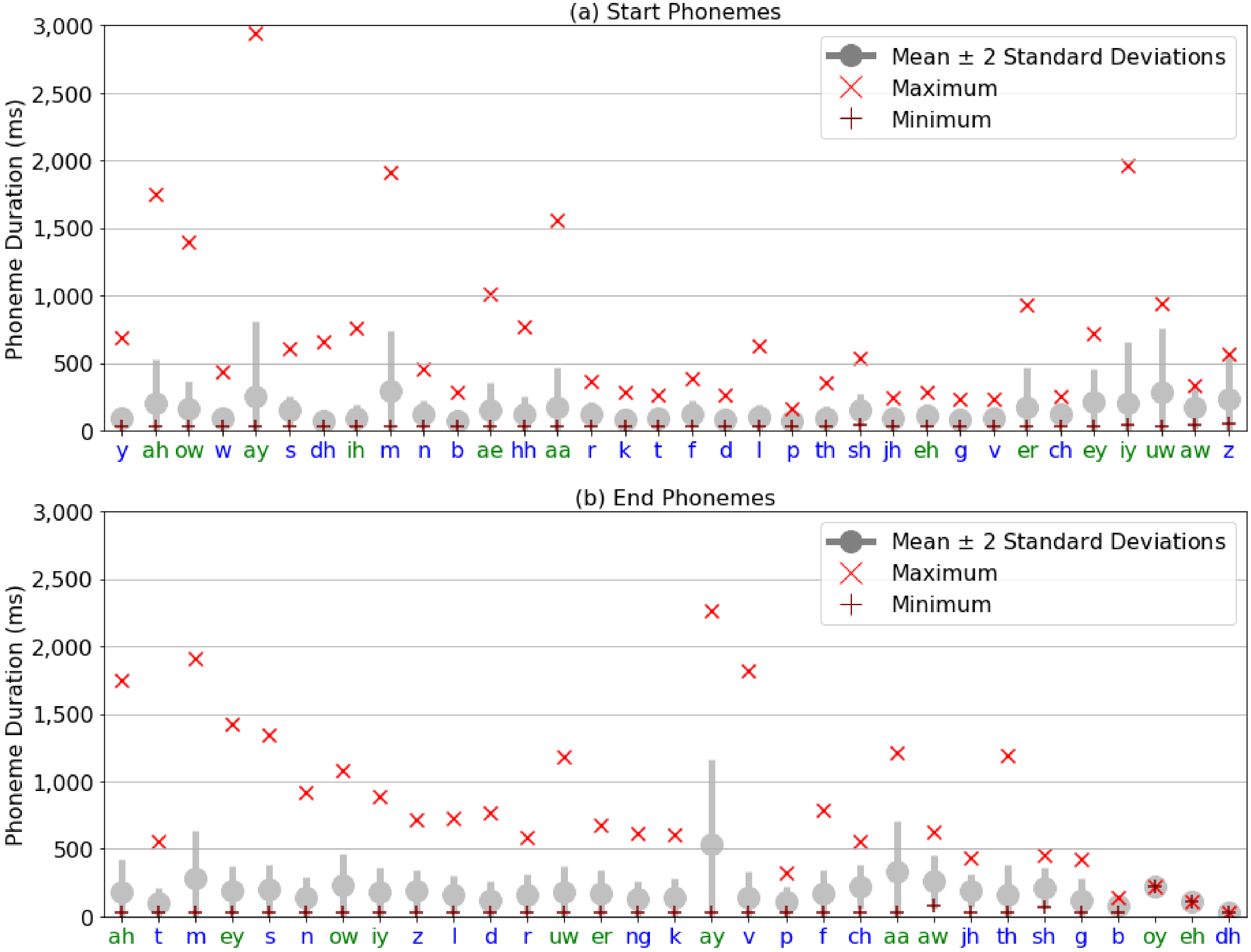
Fairly even distribution of all phonemes

Some phonemes appear at start of utterances much more often than others

Similarly for phonemes appearing at end of utterances

Phoneme Durations

Too much variation to use as indicative forgiveness collar sizes



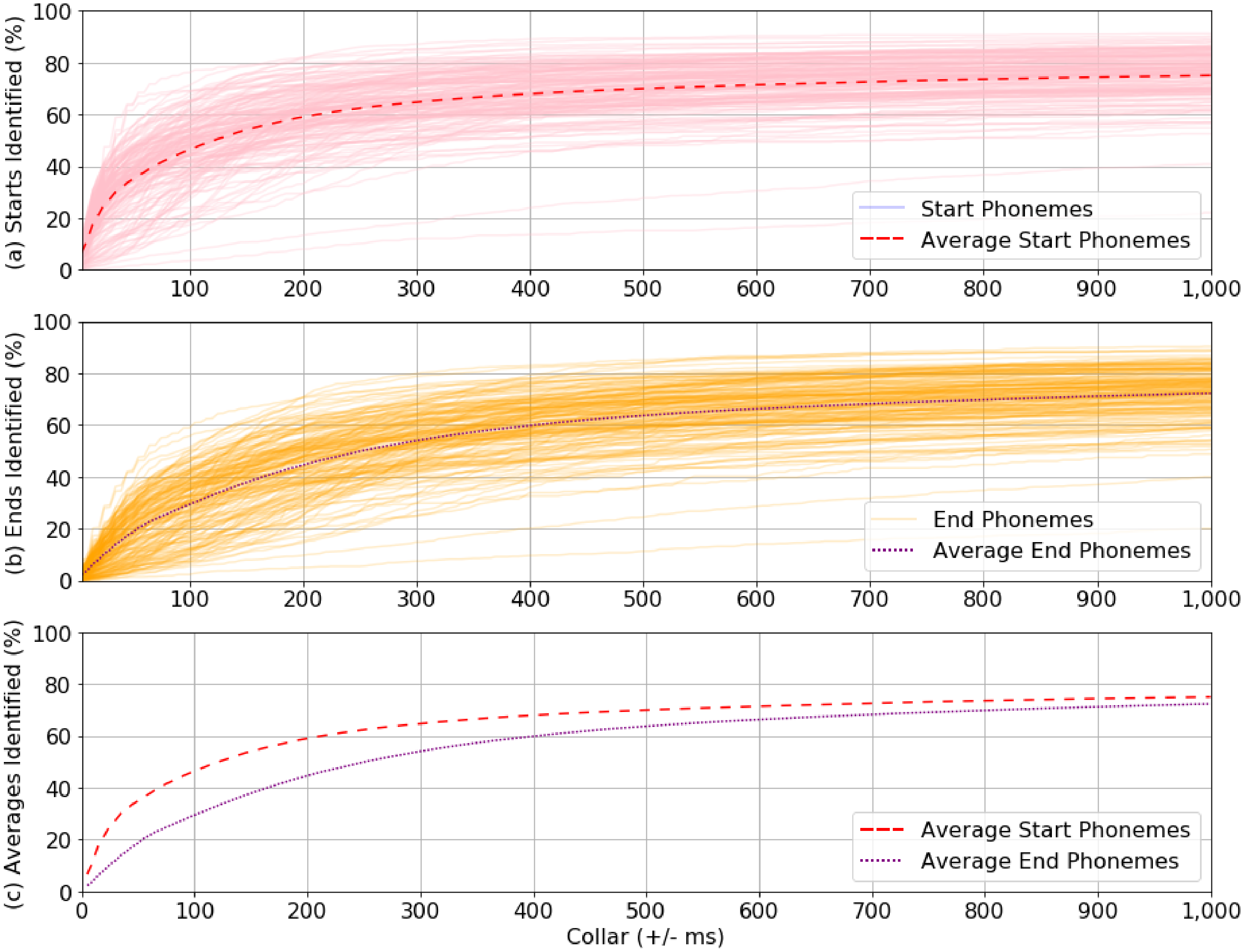
Imperial College London

Utterance Start and End Phoneme Hit Rates

All utterances assumed to be between a starting "sil" and an ending "sil"

Speaker segmentation energy based

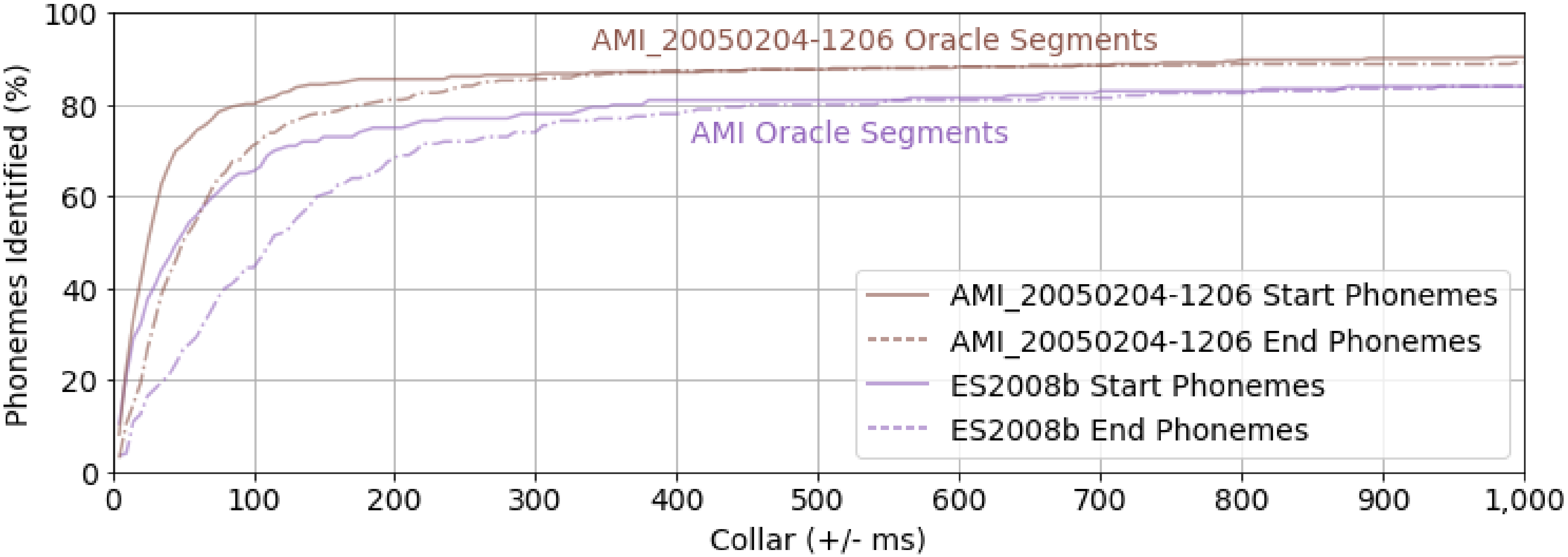
… but phoneme times generated using HTK



Start/End Phoneme Indentification Better for More Refined Ground Truth Segments

Unsurprisingly, would expect better results with more accurate segmentation

- would expect even better if phoneme identification system closely aligned with ground truth segmentation
- … but cost and effort of doing this manually is problematic

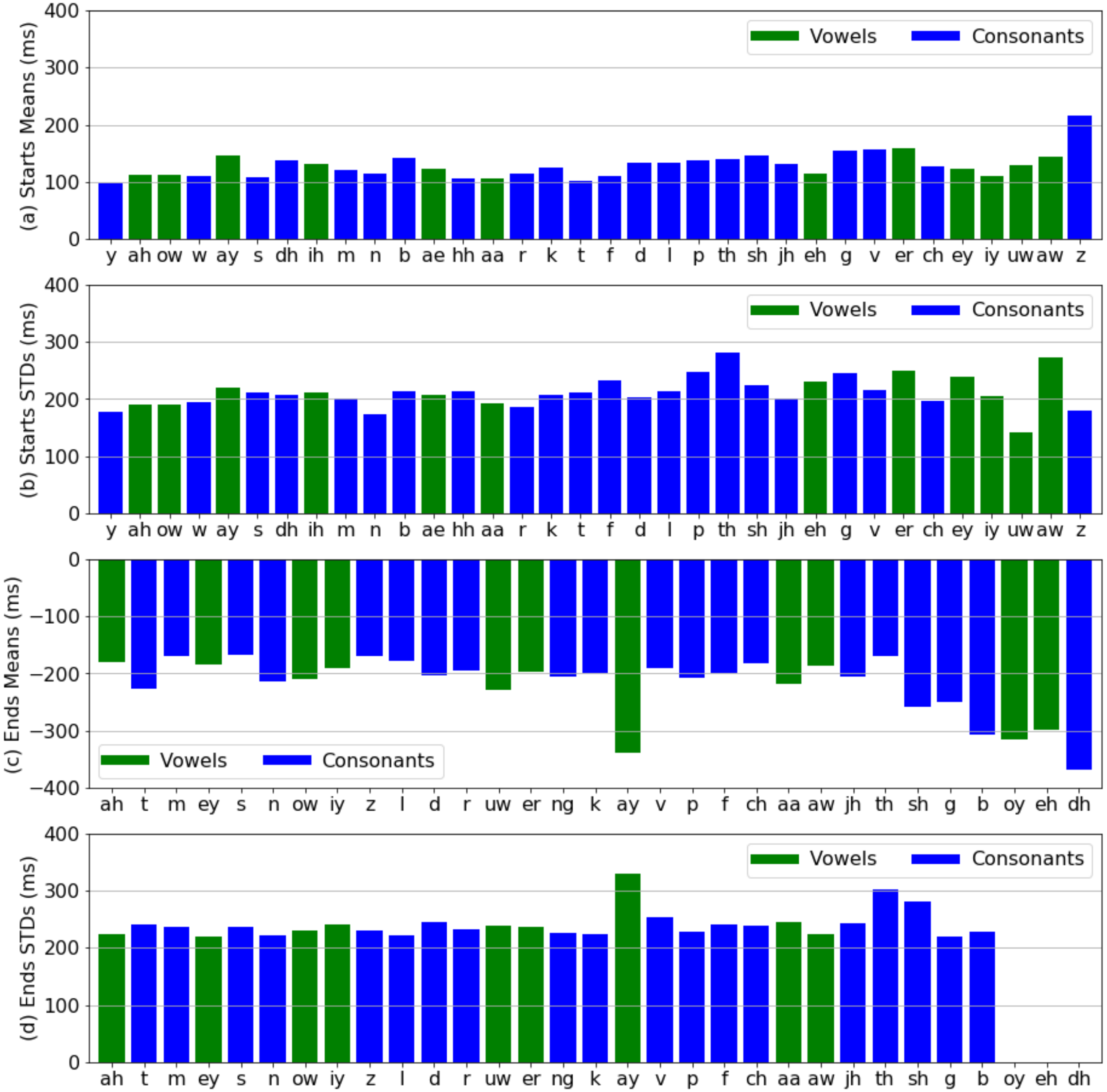


Imperial College London

Uncertainties?

Distances from ground truth utterance boundaries to start/end phonemes

- Ordered by decreasing frequency of occurrence (i.e. “y” is most common starting phoneme)
- More uncertainty for end phonemes
- Ground truth segments predict longer segments than phonemes do – need VAD/SAD that links the utterance boundaries with the phoneme boundaries
- Phoneme-dependent collars?



Imperial College
London

Conclusion

- Research shows considerable uncertainty in determining exact start and end utterance times:
  - can lead to inaccuracies in ground truth segmentation that unfairly penalize speaker diarization systems that correctly determine when utterances should start and end
- Evaluation tools that account for phonemes at utterance boundaries and whether they appear at the start or at the end of an utterance could give a better assessment of the performance of diarization systems
  - particularly useful if speaker diarization combined with speech recognition
- Next steps for DER calculations:
  - distinguish important errors from less important ones?
  - determine utterance boundaries in consistent manner with start/end phoneme times
  - introduce reliable phoneme-dependent collars?