Role Specific Language Models for Processing Neuropsychological Exams

Tuka Alhanai[1], Rhoda Au[2], and James Glass[1]
[1]MIT Computer Science and Artificial Intelligence Lab, Cambridge MA USA,
[2]Boston University School of Medicine and Public Health, Boston MA USA

1. Challenge

To screen for cognitive impairment through automatic speech processing. Easy and less invasive to record than laboratory tests and brain scans.

- **Costly**: Medical speech studies use manually generated references.
- **Generic**: Researchers model healthy speakers.

In this work, we bridge between the gap by automating data curation for medical research use.

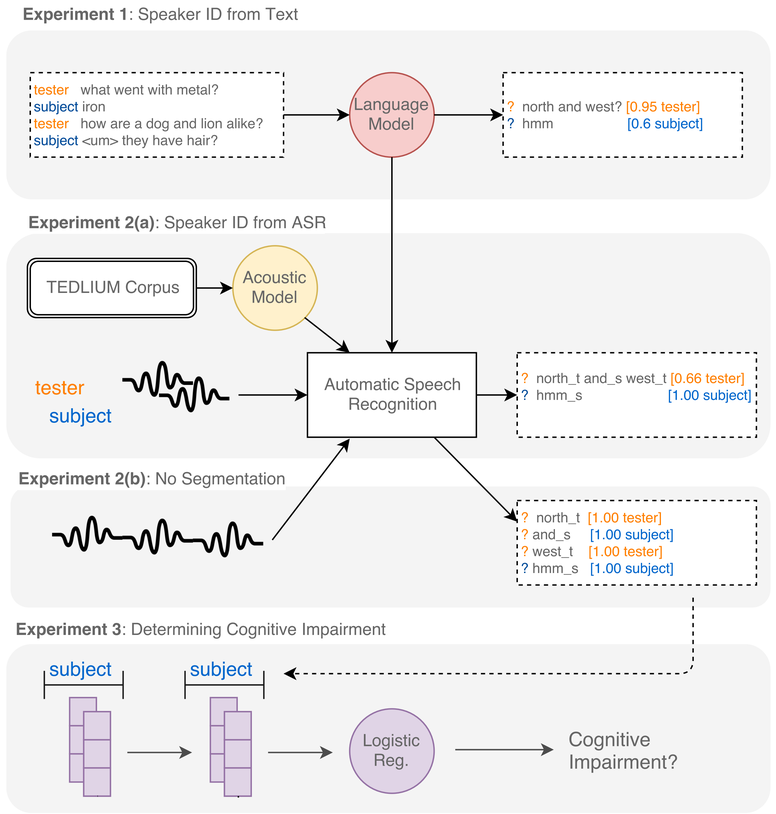
2. Speech Data

The data consists of speech recordings of neuropsychological exams that are part of the larger Framingham Heart Study:

- 92 annotated audio recordings of neuropsychological exams (~100 hours).
- Average exam is 65 minutes long, has 2,500 words, and 500 word vocabulary.
- Each recording consists of a sequence of test questions and patient responses.

Outcome of interest is cognitive impairment – 21 out of 92 subjects. Ground truth obtained from medical committee judgment.

3. Experiments



4. Setup

Exp. 1: *Features*: OOV, perplexity. *Model*: LR.
Exp. 2: *Features*: 40 filterbank + pitch. *Acoustic Model*: 6x2048 DNN trained on TEDLium Corpus.
Exp. 3: *Features*: 220 features of prosody (F0, voicing, HNR, shimmer, jitter) and energy (MFCCs, RMS energy). *Model*: regularized LR.

5. Results

	Confusion Rate	Word Error Rate	AUC
1. Text	16%	-	0.70
2. Audio	37%	81%	0.68
3. Audio + *N* seg.	.02%	81%	0.76

6. Discussion

Exp. 1: There are significant differences between speaker styles.
Exp. 2: Even with high word error rates (81%), we can diarize well (37% confusion).
Exp. 3: Possible to model cognitive impairment with noisy diarization (0.68 AUC).

- We can do even better than ground truth segmentation if we model with only 9 segments (AUC 0.76 vs. 0.70) – 150 seconds and 7% of subject's data.

7. Future Work

- Use this pipeline to process 6,000+ recordings.
- Perform population-level modeling.
- Consider underlying test being performed.
- Further evaluate weight of different segments for modeling.