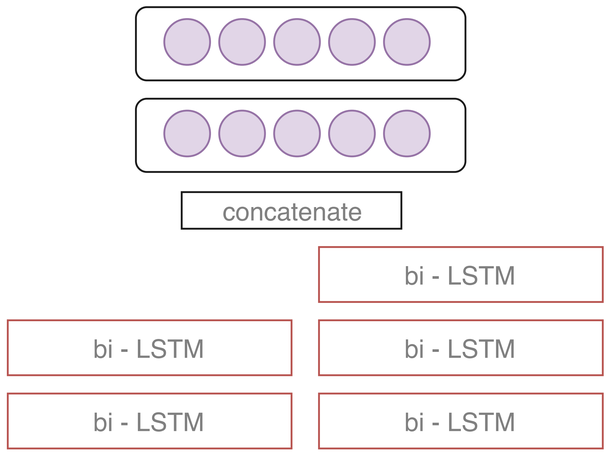
Detecting Depression with Audio/Text Sequence Modeling of Interviews

Tuka Alhanai[1], Mohammad Ghassemi[2], and James Glass[1]
[1]MIT Computer Science and Artificial Intelligence Lab, Cambridge MA USA
[2]MIT Institute for Medical Engineering and Science, Cambridge MA USA
{tuka, ghassemi, glass}@mit.edu -- talhanai.com -- github.com/talhanai

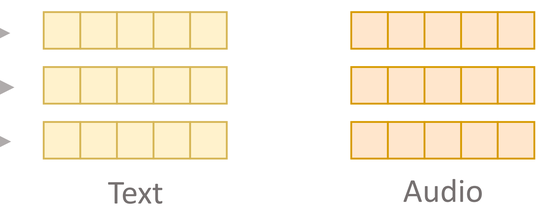
1. Objective

To screen for depression through automatic speech and language processing. Easier to record than clinical interviews. The novelty of our work:

- **Context-independent**: Our model does not depend on questions/topic of interview. (e.g. Do you have a history of depression?).
- **Feature Engineering free**: Our model uses general audio/text features. (e.g. not keyword spotting).



Q: What are some things you really like about L.A.?
A: I like the weather, I like the opportunities.
Q: How easy was it for you to get used to living in L.A.?
A: It took a minute, somewhat easy.
Q: What are some things you don't really like about L.A.?
A: Congestion.



2. Data

Corpus: DaicWOZ Corpus.
Recordings: 142 (train = 107, test = 35).
Structure:
Wizard of Oz dialogues
170 unique questions.
8,050 responses.
Outcome: binary (28 depressed).
From PHQ-8 depression screening questionnaire.

4. Setup

Audio: (x 279 features)
Higher order statistics of response segment (mean, max, min, median, standard deviation, skewness, kurtosis).
Spectral energy, prosody, and voice quality from frame-level features via COVAREP toolkit.

Text: (x 100 features)
Doc2Vec embeddings of question-response.
dim=100, min-words=3, c-win=3, epochs=50.

5. Results

Model	Features	F1	Prec.	Rec.
Baseline Approaches				
Baseline [20]	(Ensemble)	.50	.60	.43
Williamson *et al.* [6]	(Audio)	.50	/	/
Ma *et al.* [15]	(Audio)	.52	.35	1.00
Gong *et al.* [9]	(Ensemble)	.70	/	/
Williamson *et al.* [6]	(Text)	.76	/	/
†Williamson *et al.* [6]	(Text)	.84	/	/
Our Approach				
Context-free	(Audio)	.50	.71	.38
Context-free	(Text)	.59	.71	.50
Weighted	(Audio)	.67	**1.00**	.50
Weighted	(Text)	.44	**1.00**	.29
Sequence	(Audio)	.63	.71	.56
Sequence	(Text)	.67	.57	.80
Multi-modal	(Audio+Text)	**.77**	.71	**.83**

6. Discussion

1. **Sequence Modeling:**
   Highest F1 score (0.77).
   Better than baselines.
2. **Modality Inputs to Model:**
   audio = 30 sequences
   text = 7 sequences
   Depression cues exist at longer speech intervals.
3. **Weighted Model**:
   Overfits on text, better generalization with audio.

Future Work:
Decipher patterns model is capturing.
(e.g. speaking rate? Monotony?)