Using Prosodic and Lexical Information for Learning Utterance-level Behaviors in Psychotherapy

Karan Singla, Zhuohao Chen, Nikolaos Flemotomos, James Gibson, Dogan Can, David C. Atkins, Shrikanth Narayanan

USC University of Southern California

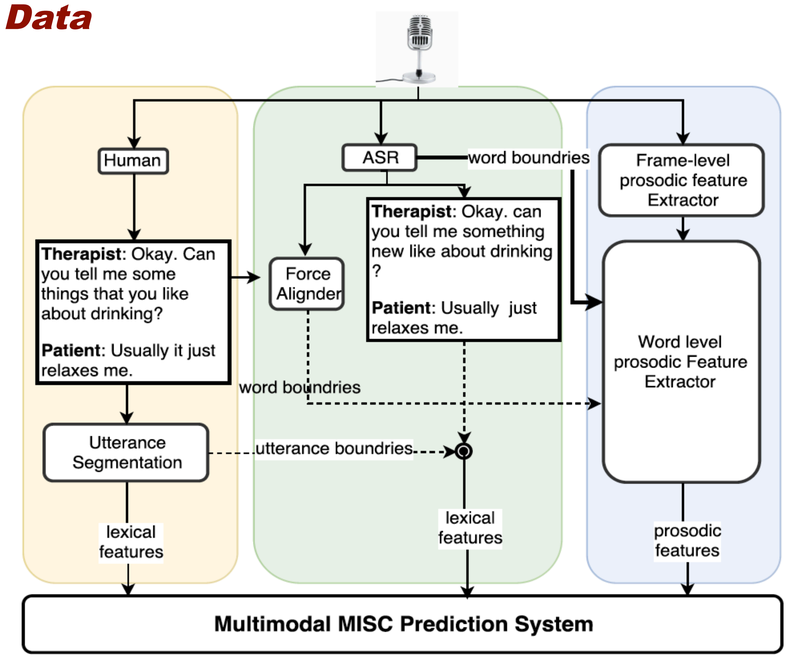
University of Washington

A multimodal system that uses both prosodic and textual information for behavior code prediction in Motivational Interviewing sessions.

Highlights

- Combined model for patient and therapist codes.
- Changes in word-level prosodic patterns can predict behavior codes.
- Prosodic information improves over text-based system for the task.
- Model gives insight into the effect of these feature streams by use of self-attention mechanism.

Data

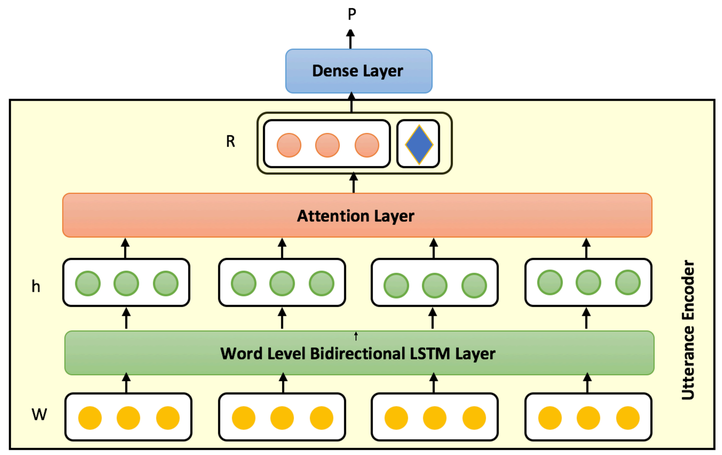


Code	Description	#Train	#Test
Therapist (T)			
REF	Reflection	6577	3456
QES	Question	6546	3348
OTH	Other	13112	7625
Total		26235	14429
Patient (P)			
FN	Follow/Neutral	22020	12229
NEG	Sustain Talk	4019	1660
POS	Change Talk	3151	1272
Total		29190	15161

Patient: I really think I can quit this time. {POS}
Therapist: So you feel confident that you can quit. {REF} What gives you that confidence? {QES}

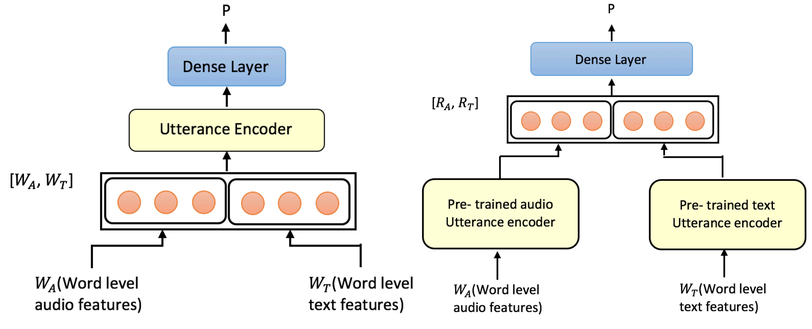
General Method

Feature Type	Prosodic					Lexicon
	Pitch	Loudness	Jitter	Pause	Word Length	
Word-level	Mean and Standard Deviation			Quantized into a 10-bit vector	Aligned Duration	Word Embedding
Dimensions	2	2	2	10	1	100
Two-fold speaker normalization: z-normalization for each audio feature for each study type; normalize each audio feature for each speaker.						



Architecture for Utterance Encoder. ◆ can be 1 or 0 for therapist and patient utterance respectively.

Multimodal Approach



Comb-WL: Word-level lexical features T and prosodic features A are word-wise concatenated to make input W before feeding it to the utterance encoder.
Comb-LF: Train utterance encoder using lexical features and a separate encoder using prosodic features, and then concatenate them before the dense layer.

Results

Features	Avg. F1 score	
	LSTM without Attention	LSTM with Attention
Text	0.54	0.57
Prosodic	0.42	042
Comb-WL	0.56	0.58
Comb-LF	0.58	0.60

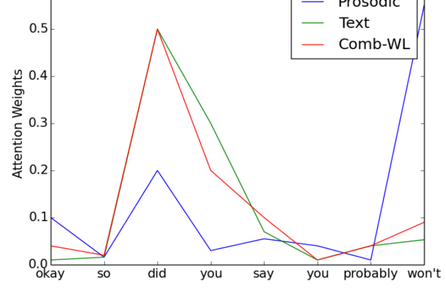
Features	Avg. F1 score
ASR text	0.47
Comb-WL	0.52
Comb-LF	0.53

- Prosodic model performs better than majority class baseline, where the avg. F1 score is 0.33.
- Multimodal approach helps in making a better prediction; Comb-LF gives the best performance.
- Results using ASR have a similar trend.

Ablation experiments where we only evaluate on Utterances > 15 words

Utterances length > 15 words	
Features	**Avg. F1 score**
Prosodic	0.48
ASR text	0.50
Comb-WL	0.54
Comb-LF	0.55

Comparison of attention weights for one question sample (QES)



- For long utterances performance of prosodic system is comparable to text-only system.
- Prosodic information at the beginning and end has greater discriminative power compared to text.

Conclusion

- Using prosodic features in addition to lexical features aids in the prediction of utterance-level behaviors in psychotherapy.
- Attention layer helps improve the performance.
- Comb-LF outperforms other models with both human transcribed text and automatically transcribed text.
- Encoder architecture efficiently exploits word-level prosodic variation.