A MULTIMODAL ANALYSIS OF SYNCHRONY DURING DYADIC INTERACTION USING A METRIC BASED ON SEQUENTIAL PATTERN MINING

Shanghai, China
ICASSP·2016
The 41st IEEE International Conference on Acoustics, Speech and Signal Processing, 20-25 March 2016

Anil Jakkam and Carlos Busso

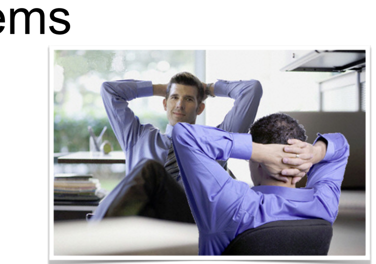
Multimodal Signal Processing (MSP) Laboratory
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas

MOTIVATION

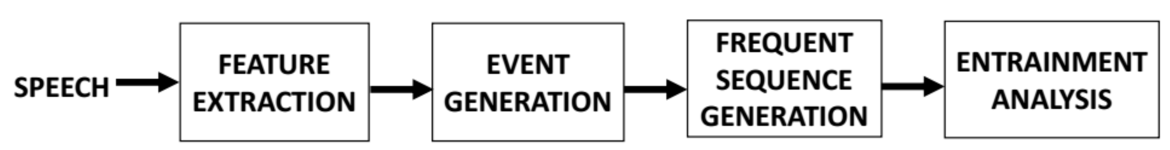
Background:

- Adaptation of interlocutors in human conversations is called synchrony, entrainment, or mimicry
- Synchrony has been studied well within single modality
  - synchrony across modalities is an open question
- Measurement and quantification of synchrony could:
  - improve existing spoken dialogue systems
  - Improve emotion recognition systems

Proposed Solution:



- We use sequential pattern mining to study the role of synchrony in dyadic conversations
- We study synchrony at a turn level considering the acoustic and text modalities on the Fisher's Corpus



Framework

Fisher's corpus

- Dyadic interactions, with randomly assigned speakers
- 90 sessions (30 each for training validation and testing)

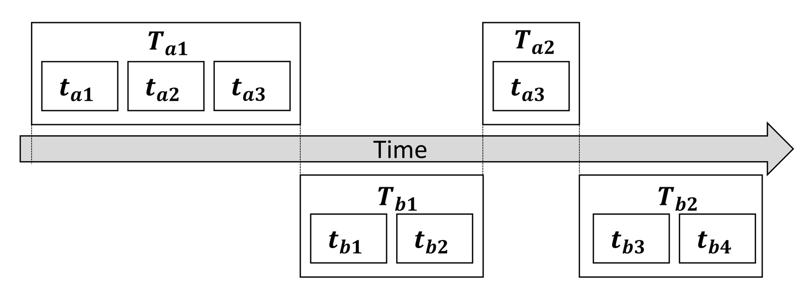
Sequential Pattern Mining

- Find frequent sequence of events
- Definitions:
  - event (e_k): relevant observations
  - itemset (i_k): unordered list of events (e_1 e_2 ... e_n)
  - sequence: ordered list of itemsets <i1,i2,...,in>
  - support: # of data sequences containing a given sequence

- Example:

Seq. #	Sequence
1	<(a)(b)>
2	<(ac),(b)>
3	<(abc)(ab)>
4	<(a)(ab)>

  - Seq. <(a),(b)> has support 4
- We use two itemsets corresponding to speaking turns
  - <(speaker 1), (speaker 2)>
- Definition of speaking turn:



- All events are considered simultaneous within a turn

- Events from audio and text
  - Intensity (4 events)
  - F0 (4 events)
  - Disfluency (4 events)
  - Duration (6 events)
  - laughter (1 event)
  - ToBI (20 events)

		Events
Energy	1	High intensity
	2	Least min intensity
	3	Highest max intensity
	4	Highest range intensity
F0 Contour	5	Highest F0
	6	Least min F0
	7	Highest max F0
	8	Highest range F0
Disfluencies	9	Disfluency-Fillers
	10	Disfluency-Discourse marker
	11	Disfluency-Editing term
	12	Disfluency-Repetition
Duration	13	Low Turn Duration
	14	High Turn Duration
	15	Low phoneme rate
	16	High Phoneme Rate
	17	Low word rate
	18	High Word Rate
	19	Laughter

		Events
ToBI	20	# H*: high pitch accent
	21	# L*: low pitch accent
	22	# L+H*: bitonal pitch accent with low tone followed by high tone prominence
	23	# L*+H: bitonal pitch accent with low tone prominence followed by high tone
	24	# !H*: downstepped high pitch accent
	25	# L+!H*: bitonal pitch accent with low tone followed by a downstepped high tone prominence
	26	# L*+!H: bitonal pitch accent with low tone prominence followed by downstepped high tone
	27	# H+!H*: bitonal pitch accent with high tone followed by downstepped high prominence
	28	# L-L%: low phrase accent, low boundary tone
	29	# H-H%: high phrase accent, high boundary tone
	30	# L-H%: low phrase accent, high boundary tone
	31	# H-L%: high phrase accent, low boundary tone
	32	# !H-L%: downstepped high phrase accent, low boundary tone

Finding Relevant Sequences

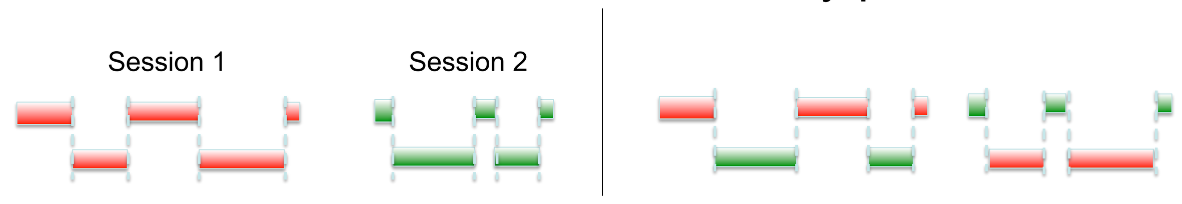
Selection of Relevant Sequences

- SPADE algorithm (SPMF)
- Step 1: Discovers frequent sequences in training set
  - remove sequences with low support
  - contains over 1000 sequences

Seq. #	Sequence	SUP
1	<(9),(9)>	0.185
2	<(14),(9)>	0.183
3	<(9),(14)>	0.149
4	<(36)(9)>	0.144
5	<(10)(9)>	0.138
6	<(14,36)(9)>	0.133
7	<(1)(9)>	0.128
8	<(14),(14)>	0.124
9	<(9)(10)>	0.122
10	<(14)(10>	0.113

  - Sequences may not inform about synchrony (e.g., <(9),(9)>)

- Step 2: Discovers sequences relevant to synchrony
  - We use the validation set
  - Paired condition vs. Randomly paired condition

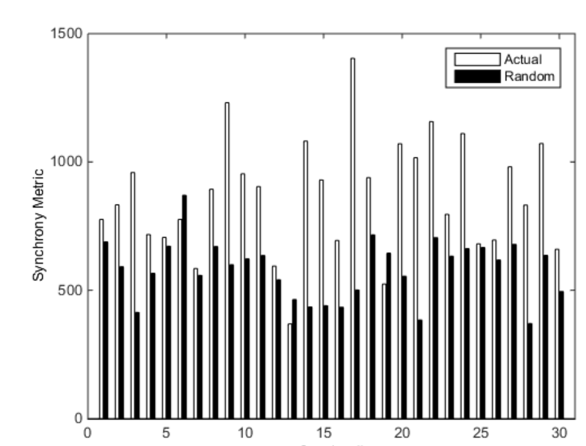


  - Estimate support of sequences in master list
  - Compute ratio of their support and select top 100 sequences

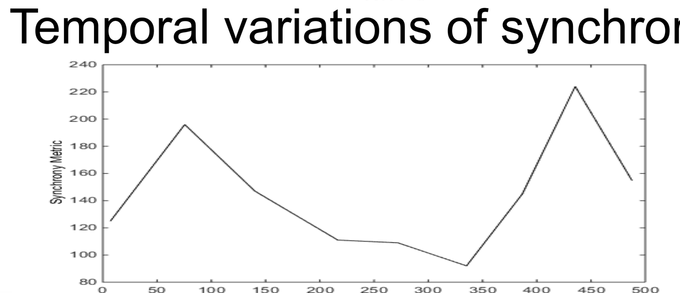
High turn duration (14) and **downstepped high pitch accent (24)** of one speaker, triggers a **low phoneme (15) and word rate (17)** on the other speaker.

Seq. #	Sequence
1	<(14,24,36)(15,17)>
2	<(24,36)(15,17)>
3	<(14,24)(15,17)>
4	<(14,24,36)(17)>
5	<(1,14,36,39)(15)>
6	<(1,36,39)(15)>
7	<(1,24,36)(15)>
8	<(1,14,24,36)(15)>
9	<(1,14,36,39)(17)>
10	<(1,36,39)(17)>

- Step 3: Define a metric of synchrony
  - We use the testing set
  - Metric: sum of the support of the top 100 sequences for a given session
  - Paired condition > Randomly paired condition (27 out of 30 sessions)



  - Temporal variations of synchrony



DISCUSSION

Conclusions:

- This framework captures the local interplay of multiple modalities that lead to synchrony
- Sequential pattern mining is a fast and efficient way to discover frequent sequences
- We developed a metric that effectively represents synchrony

Future Work:

- Use the sequential features to classify engagement, depression, or empathy.
- Incorporate a variable window, rather than just considering the adjacent turns
- Extension to multiparty interaction
- Using other modalities

Acknowledgment: This study was funded by NSF (grant IIS1217104) and a NSF CAREER award IIS-1453781