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

Anil Jakkam and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory
The University of Texas at Dallas, Richardson TX 75080, USA
akj140230@utdallas.edu, busso@utdallas.edu

ABSTRACT

In human-human interaction, people tend to adapt to each other as the conversation progresses, mirroring their intonation, speech rate, fundamental frequency, word selection, hand gestures, and head movements. This phenomenon is known as synchrony, convergence, entrainment, and adaptation. Recent studies have investigated this phenomenon at different dimensions and levels for single modalities. However, the interplay between modalities at a local level to study synchrony between conversational partners is an open question. This paper studies synchrony using a multimodal approach based on sequential pattern mining in dyadic conversations. This analysis deals with both acoustic and text-based features at a local level. The proposed data-driven framework identifies frequent sequences containing events from multiple modalities that can quantify the synchrony between conversational partners (e.g., a speaker reduces speech rate when the other utters disfluencies). The evaluation relies on 90 sessions from the Fishers corpus, which comprises telephone conversations between two people. We develop a multimodal metric to quantify synchrony between conversational partners using this framework. We report initial results on this metric by comparing actual dyadic conversations with sessions artificially created by randomly pairing the speakers.

Index Terms— Human-Human Interaction; Synchrony; Entrainment; Convergence; Multimodal

1. INTRODUCTION

The adaptation of the interlocutors to one another in human conversations has been long established in early research in the field of psychology and communication science [1]. This adaptation of the speakers has been referred to with several terms such as synchrony, entrainment, mimicry, convergence, alignment, accommodation, reciprocity and mirroring. Early research in this area has focused on the evidence and measurement of entrainment in various modalities and at different levels. The studies have considered and reported evidences of synchrony in acoustic, lexical and visual modalities at local and global levels [2–8].

Previous studies have explored evidences of synchrony within a single modality (e.g., when one speaker increases his/her speech rate, the other responds by speaking faster). However, human communication is inherently multimodal. We hypothesize that synchrony is a broader phenomenon where "events" happening in one modality (e.g., producing disfluencies) affects events produced by the interlocutor in the same, or different modality (e.g., decreasing speaking rate). Identifying patterns across modalities can improve our understanding about social communication. This paper explores a data-driven framework to identify sequences of multimodal events produced by speakers that appear during natural human interactions.

This paper explores the use of sequential pattern mining to study the role of synchrony in dyadic conversations. We use prosodic and text-based features to study synchrony at the turn level. The concept of sequential patterns comes from the field of data mining, and was first introduced by Agarwal et al. [9] for market basket analysis. The framework has been used in various applications where sequence of patterns need to be identified such as biological protein sequences and website click streams [10]. It is a suitable framework to identify sequence of events across modalities that are produced by speakers during the course of natural interactions. Unlike previous work, this framework allows us to extract the local interplay of multiple modalities that lead to synchrony. This framework has been used before in the area of multimodal signal processing for modeling user experience in gaming [11]. The count of the sequential patterns were used as features, for classifying the users affect in a Maze-Ball game. This is the first time that this work is used in the context of synchrony.

2. RELATED WORK

Nenkova et al. [2] explored entrainment at the linguistic level, concluding that the use of high-frequency words was significantly correlated with task success. It was also observed that higher degrees of entrainment are associated with more overlaps and fewer interruptions during the interaction. Lee et al. [3] quantified entrainment at the acoustic level using the information from the fundamental frequency and energy in speech (Pearson correlation, mutual information, and mean of spectral coherence). They also successfully used these quantized entrainment measures to classify positive and negative affects with an accuracy of 76%. Heldner et al. [4] studied local entrainment in the fundamental frequency, demonstrating that the F0 contour during backchannels was more similar to the immediately preceding utterances of the opposite speaker than during non-backchannels. Levitan and Hirschberg [5] measured and quantified entrainment at multiple levels and dimensions. They explored entrainment at both the turn and session level and also studied its effect in four acoustic dimensions - energy, fundamental frequency, speaking rate and voice quality. Also, three different views were considered, namely, proximity, convergence and synchrony. De Looze et al. [7] studied the dynamics of prosodic accommodation within and across dyadic telephonic conversations. Scherer et al. [12] studied the effect of accommodation on depression severity assessment during interviews. Gravano et al. [6] provided a measure of prosodic entrainment capturing *backward mimicry*, where the F0 contour used by a speaker, was previously used by the interlocutor, and *forward influence*, where the F0 contour used by a speaker is used by the interlocutor in the following turns. They compare this metric with the level of engagement. Xiao et al. [8] derived a measure for entrainment using speech rate, investigating its relation with empathy.

This study was funded by National Science Foundation (NSF) grant IIS 1217104 and a NSF CAREER award IIS-1453781.

Table 1. Sequence Database - Example

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

The measurement and quantification of entrainment could be used to improve the existing spoken dialogue systems, by incorporating the complex dynamics involved in social interaction [7]. Existing emotion recognition systems could be improved by leveraging emotional entrainment. For example, Lee et al. [13] used speech based features to model the mutual influence, which was then used to recognize the emotion of the participants improving the classification accuracy. Mariooryad and Busso [14] used acoustic and visual features, utilizing cross-modality and cross-speaker information to improve the performance of an emotion recognition system. Bell et al. [15] reported entrainment between a user and an interface. By reducing the speech rate of the interface, they indirectly influenced the speech rate of the users, increasing the performance of an *automatic speech recognition* (ASR) system.

All these studies explore synchrony only within a single modality. They do not address the problem of synchrony across modalities, which is the focus of this paper.

3. THE FISHER'S CORPUS

This study relies on the Fisher's corpus, which is a large database of conversational telephone speech, collected by the *Linguistic Data Consortium* (LDC) [16]. In every session (or call) the speakers were randomly assigned topics from a list. Since the speakers are randomly assigned for every telephone call, it is expected that the speakers naturally tend to adapt to one another as the conversation progresses, depicting some form of synchrony. This is the reason for choosing the Fisher's corpus for our study on synchrony. Each conversation lasted for about 10 minutes. This study uses the speech files and the transcriptions of the conversations to derive the acoustic and text-based features. We use forced alignment to determine the phoneme and word alignments.

We only consider the first 90 sessions from the Fishers corpus which are divided into three partitions with 30 sessions each: training, validation and testing. We use the training set to create a master list with all the frequent sequences (Sec. 5.2). We use the validation set to select the best sequences that are indicative of synchrony (Sec. 5.3). We use the testing set to evaluate the selected sequences on different recordings (Sec. 6).

4. SEQUENTIAL PATTERN MINING

This section briefly explains the concept of sequential pattern mining and defines its terminology. For a given database of sequences, sequential pattern mining finds all the *sequential patterns* that occur more frequently than a given support (i.e., a user defined threshold). An *event* e_k is an observation that we are interested, which in our study corresponds to cues such as disfluencies, and an increase in F0 value. A set of events forms an *itemset* i_k, which is an unordered list of events of the form $(e_1 e_2 ... e_n)$. Events in an itemset have no temporal order and are assumed to occur simultaneously within the itemset. A sequence is an ordered list of itemsets of the form $< i_1 i_2 ... i_n >$, where each i_k represents an itemset. The support for a sequence is the fraction of the total data sequences that contain this particular sequence.

Consider the example of the sequence database shown in Table 1. Here, a, b, c, d are events, events within parenthesis forms the itemsets (e.g., (a, b)), and each row in the table is a sequence

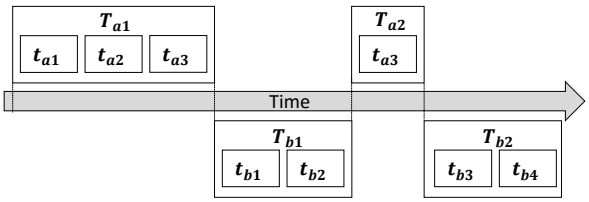

Fig. 1. Definition of speaking turns. The speaking turn is always followed by a speaking turn of the interlocutor.

(e.g., $<(a)(b)>$; a is an event in an itemset, and b is an event in the next itemset). Suppose, the minimum support is 2, then $<(a)(b)>$ would be a frequent sequence, since it has a support of 4. Notice that $<(a)(b)>$ is a subsequence of all the four sequences $<(\mathbf{a})(\mathbf{b})>$, $<(\mathbf{a}c)(\mathbf{b})>$, $<(\mathbf{a}bc)(a\mathbf{b})>$ and $<(\mathbf{a})(a\mathbf{b})>$, so it has support of 4.

We adapt this framework to our problem by considering events such as high intensity, high F0 contour, laughter, or high turn duration. First, we define speaking turns as the segments starting when one subject begins to speak and finishing when the next subject begins to speak. These segments can include multiple sentences or phases (see Fig. 1 where T_{a1} and T_{a2} represent speaking turns that have multiple sentences t_{a1} and t_{a2}). With this definition, speaking turns alternate between speakers. Second, we extract all the events observed in the segment, defining an itemset for each speaking turn. We create the sequence database by adding consecutive pairs of itemsets, i.e. an itemset of the turn of one speaker followed by the itemset of the immediate turn of the other speaker in a conversation. These pairs of itemsets represent important local information capturing sequence of events across speakers (e.g., one speaker increases the intensity, the other speaker laughs).

As the list of sequences and the subsequences is huge, this framework uses sequential pattern mining to discover the sequences that frequently happen in the data. This is a novel, data driven, multimodal framework that fuses local information from the acoustic and text-based features by building the sequence databases. We employed the SPADE algorithm [17] to extract the frequent sequences. The SPADE algorithm decomposes the original problem of finding the frequent sequences into smaller sub-problems, which can be independently solved using efficient lattice search techniques and simple join operations. We use the implementation of SPMF, an open source data mining Java library [18].

5. METHODOLOGY

This section presents the steps involved in mining the frequent sequences from the multimodal data (Fig.2). We extract the features, generate events for each speaking turn, build a sequence database, and discover the frequent sequences based on the sequential pattern mining approach. Finally, we identify the best sequence pairs which we expect to describe synchrony.

5.1. Defining the Events

The proposed framework is flexible, allowing us to consider events across multiple modalities. We have incorporated a wide range of events, so that the framework can discover the frequent sequences. This study evaluates the approach using prosodic and text-based features. From speech, we estimate the intensity and fundamental frequency using the OpenSMILE toolkit [19] over frames of 25ms with

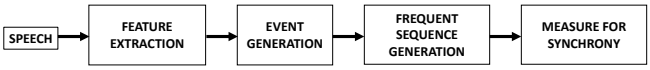

Fig. 2. Block Diagram of the framework

Table 2. List of events used in this study. The events are automatically extracted from the speech and transcription of the recordings.

#	Event	#	Event	#	Event	#	Event	#	Event
1	High intensity	9	Disfluency-Fillers	17	Low word rate	25	# L+!H*: bitonal pitch accent with low tone followed by a downstepped high tone prominence	33	# H-: high phrase accent
2	Least min intensity	10	Disfluency-Discourse marker	18	High Word Rate	26	# L*+!H: bitonal pitch accent with low tone prominence followed by downstepped high tone	34	# L-: low phrase accent
3	Highest max intensity	11	Disfluency-Editing term	19	Laughter	27	# H+!H*: bitonal pitch accent with high tone followed by downstepped high prominence	35	# !H-: downstepped high phrase accent
4	Highest range intensity	12	Disfluency-Repetition	20	# H*: high pitch accent	28	# L-L%: low phrase accent, low boundary tone	36	Break Tier 1
5	Highest F0	13	Low Turn Duration	21	# L*: low pitch accent	29	# H-H%: high phrase accent, high boundary tone	37	Break Tier 2
6	Least min F0	14	High Turn Duration	22	# L+H*: bitonal pitch accent with low tone followed by high tone prominence	30	# L-H%: low phrase accent, high boundary tone	38	Break Tier 3
7	Highest max F0	15	Low phoneme rate	23	# L*+H: bitonal pitch accent with low tone prominence followed by high tone	31	# H-L%: high phrase accent, low boundary tone	39	Break Tier 4
8	Highest range F0	16	High Phoneme Rate	24	# !H*: downstepped high pitch accent	32	# !H-L%: downstepped high phrase accent, low boundary tone		

Table 3. Master List - Top 10 sequences with the highest support.

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

a step size of 10ms. We estimate the minimum, maximum and range of the intensity and fundamental frequency for each speaking turn. The text-based events describe disfluency, turn duration, phoneme rate, word rate and laughs. We extract these features from the transcriptions of the recordings and the phoneme and word alignments.

For simplicity, most of the events correspond to either lower or higher values of certain features (e.g., high fundamental frequency). For this purpose, we estimate the distributions of the features per speaker, computing first and third quartile. Low values correspond to features below the first quartile and high values correspond to features above the third quartile. There are 39 events in total listed in Table 2, which are described next.

5.1.1. Acoustic Events

An event of high intensity (1) occurs within a turn, when the number of intensity peaks is greater than a fixed threshold. First, we estimate the distribution of intensity values across all frames. We select intensity peaks by locating segments in which their intensity values exceed the upper third quartile in the intensity distribution. Then, we count the number of intensity peaks per speaker turn. An event is detected in a turn if its number of intensity peaks is higher than the value associated with the third quartile of number of peaks per speaking turn. The thresholds are separately determined for each speaker. For events involving the functionals - minimum, maximum and range, we first extract the functionals for each turn. For a given speaker, we determine a threshold based on either the first or the third quartile over the set of the functionals across all turns. The event of least minimum intensity (2) was determined based on the first quartile threshold. The events the highest maximum intensity (3) and the highest range (4) were obtained based on the third quartile of their respective values. We also estimate events from the fundamental frequency following a similar approach. This study considers the following events: high F0 (5), the least minimum (6), the highest maximum (7) and the highest range (8) of F0 contour values.

We define 21 events related to the *Tone and Break Index* (ToBI). We estimate ToBI labels using the toolkit AuToBI [20]. In tone tier, the subtle changes in the prosody information are captured and represented with the symbols like H* for high pitch accent (20), L* for low pitch (21). We include other ToBI labels (22-35), as shown in Table 2. The break tier (36-39) represents the amount of disjuncture between words and has a value from '0' to '4'. An index of '1' indicates a typical word boundary, whereas '4' indicates an intonational phrase boundary. We count the number of tone and break indices over turns, defining events based on the third quartile threshold.

5.1.2. Text Events:

From the transcriptions, we estimate 11 events. First, we estimate disfluencies, where we consider four types: fillers (9) such as 'uh' and 'um', discourse markers (10) such as 'well' and 'you know', editing terms (11) like 'I mean' and 'sorry', and repetitions (12) [21]. The presence of each of these disfluencies in a turn is labeled as an event. The turn duration, phoneme rate and word rate are obtained from the forced alignment files for each turn. The number of phonemes and words are counted and divided by the turn duration to obtain the phoneme and word rates, respectively. If the turn duration, phoneme rate and word rate for a turn are greater than the third quartile, they are labeled as high turn duration (14), high phoneme rate (16) and high word rate (18) events, respectively. If they are less than the first quartile, they are labeled as low turn duration (13), low phoneme rate (15) and low word rate (17) events, respectively. Laughter (19) is a miscellaneous event obtained from transcriptions.

5.2. Frequent Sequence Generation

We use the training set to identify frequent sequences. We build the sequence database by populating each row with the itemsets of consecutive turns of both speakers in a session. Itemsets are generated for each turn by listing all the events in them. As we focus on synchrony at the local level, we limit the size of each sequence in the database to two itemsets. However, an itemset can contain any number of events, which are assumed to simultaneously occur during the speaking turn. Therefore, every sequence in the database contains two itemsets, one from each speaker. The frequent sequences with a single itemset are removed, as they carry information related to only one speaker, and, hence, are irrelevant to this study.

We generated a database for each session in the training set, keeping the frequent sequences with a minimum support of 5%. This support is necessary to limit the number of frequent sequences. With this support, the smallest and largest number of total sequences for the sessions in the training set are 537 and 62,736, respectively. We create a master list with 135,123 sequences by listing all the unique frequent sequences across training sessions. Some of these sequences do not appear in other sessions. Therefore, we estimate the average support the sequences over all the training sessions, set-

Table 4. Best Sequences - Top 10 sequences

Seq	Sequence	SUP	Seq	Sequence	SUP
1	<(14,24,36)(15,17)>	0.020	6	<(1,36,39)(15)>	0.021
2	<(24,36)(15,17)>	0.020	7	<(1,24,36)(15)>	0.021
3	<(14,24)(15,17)>	0.021	8	<(1,14,24,36)(15)>	0.021
4	<(14,24,36)(17)>	0.028	9	<(1,14,36,39)(17)>	0.022
5	<(1,14,36,39)(15)>	0.02	10	<(1,36,39)(17)>	0.022

ting a minimum support of 2%. This second threshold reduces the number of sequences to 1,133. Table 3 lists the top 10 sequences. For example, the first sequence indicates that after a speaker produces a filler (event #9), the other speaker will also have a filler.

5.3. Selection of Relevant Sequences

The master list contains over 1,000 frequent sequences. Some of them may be due to events that occur very often (e.g., fillers are very common). We are not interested in these events. Instead, we want to identify events associated with synchrony. For this purpose, we follow an approach commonly used by related work on synchrony. It consists of comparing results obtained from original recordings (paired condition) with recordings where we randomly pair speaking turns from different sessions (unpaired condition) [5] [8]. Events that happens only based on chances, or because they are very common events will also appears in randomly paired speaking turns. Notice that we respect the temporal order of the speaking turns. We just randomly replace the recordings from one speaker from one session for the recordings of another speaker from a different session. This analysis is conducted over the validation set, which is independent of the training set used to create the master list.

We generate ten random sequence databases for each session. A random sequence database is generated by randomly pairing one of the speaker's itemsets with the itemsets of a speaker from another session. Frequent sequences from the master list are extracted for both the actual as well as the random sequence databases. Based on the master list, the support of a sequence in each of the sessions is calculated. Like in the previous case, we separately take an average of the supports over all the sessions for the paired and unpaired conditions. Then, we compare their supports by taking their ratio. A large number indicates that a given sequence appears very often in the paired condition, but not as often in the unpaired condition. These are the sequences that we expect to be indicative of synchrony between speakers. Arbitrarily, we only keep the top 100 sequences with the highest ratio. Table 4 shows the top 10 best sequences, listed based on their ratio.

6. EXPERIMENTAL EVALUATION

We evaluate this framework with the testing set. While there are many options to use this framework, we consider a metric created by adding all the occurrences of the best 100 sequences during a conversation. We compute this metric over the actual sessions in the testing set. For illustration, we also estimate this metric for recordings with randomly paired conditions (same approach presented in Sec. 5.3). Figure 3 shows the number of these sequences that appears during the conversation for paired (gray bars) and unpaired (black bars) conditions for the 30 testing sessions. In 27 out of 30 sessions, the sessions with paired conditions have significantly higher number of selected sequences than the sessions with unpaired condition ($p<0.001$ – z test for two population means). This result shows that this measure effectively represents local synchrony between the speakers in dyadic conversations.

Table 4 shows that the events from the text-based features dominates the best sequences. Events such as low phoneme rate and high

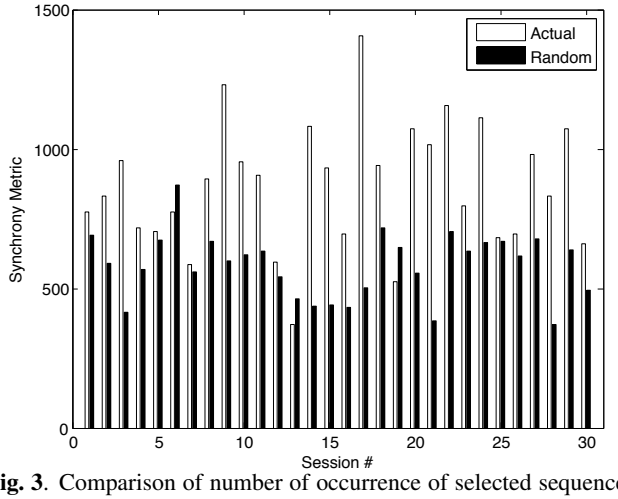

Fig. 3. Comparison of number of occurrence of selected sequences for paired and unpaired session.

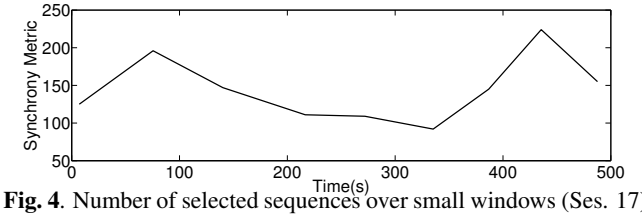

Fig. 4. Number of selected sequences over small windows (Ses. 17).

turn duration occur in majority of the best sequences. Consider sequence # 3 which says that a high turn duration and a downstepped high pitch accent of one speaker, triggers a low phoneme and word rate on the other speaker. This means that if one speaker is speaking slower, then the other person adapts to produce low phoneme and word rates, which clearly is a manifestation of synchrony. This metric is a representative measure of the rich local multimodal information present in the data. It presents tremendous potential to study other mental and cognitive conditions such as depression, dominance and empathy, which are related to synchrony.

Figure 4 shows the temporal variations of the metric for session 17, averaged over 15 speaking turn windows. This session has the highest number of selected sequences among all the sessions in the testing set. It is clear that synchrony is a local phenomenon, which varies dynamically throughout the conversation. This result agrees with the work by De Looze et al. [7], which showed that prosodic accommodation varies dynamically in a conversation.

7. CONCLUSIONS AND FUTURE WORK

We proposed a framework to capture the local interplay of multiple modalities in dyadic conversations that lead to synchrony. We transformed the analysis of synchrony to a pattern mining problem, thereby leveraging the vast potential of this domain. The sequential pattern mining provides us with a fast and efficient way to discover the frequent sequences. Furthermore, we have developed a metric using this framework which can represent synchrony. This is just a starting point in the direction of multimodal synchrony analysis. As a future work, it would be interesting to look at how these sequences can be used as features in classification of engagement in conversations, or in investigating their role in depression or empathy analysis. Also, we will incorporate a variable window, rather than just considering the adjacent turns. We will also consider other modalities and events, and potential extensions of the framework to multi-party interaction.

8. REFERENCES

[1] H. Giles, J. Coupland, and N. Coupland, *Contexts of Accommodation: Developments in Applied Sociolinguistics*, Cambridge University Press, New York, NY, USA, September 1991.

[2] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, Columbus, OH, USA, June 2008, pp. 169–172.

[3] C.-C Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P.G. Georgiou, and S.S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 793–796.

[4] M. Heldner, J. Edlund, and J. Hirschberg, "Pitch similarity in the vicinity of backchannels," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 3054–3057.

[5] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *12th Annual Conference of the International Speech Communication Association (Interspeech'2011)*, Florence, Italy, August 2011, pp. 3081–3084.

[6] A. Gravano, S. Beňuš, R. Levitan, and J. Hirschberg, "Backward mimicry and forward influence in prosodic contour choice in standard american english," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 1839–1843.

[7] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction," *Speech Communication*, vol. 58, pp. 11–34, March 2014.

[8] B. Xiao, Z.E. Imel, D. C. Atkins, P. G. Georgiou, and S.S. Narayanan, "Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 2489–2493.

[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD Record*, vol. 22, no. 2, pp. 207–216, June 1993.

[10] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data Mining and Knowledge Discovery*, vol. 15, no. 1, pp. 55–86, August 2007.

[11] H.P. Martínez and G. N. Yannakakis, "Mining multimodal sequential patterns: A case study on affect detection," in *International conference on multimodal interfaces (ICMI 2011)*, Alicante, Spain, November 2011, pp. 3–10.

[12] S. Scherer, Z. Hammal, Y. Yang, L.P. Morency, and J. F. Cohn, "Dyadic behavior analysis in depression severity assessment interviews," in *International conference on multimodal interaction (ICMI 2014)*, Istanbul, Turkey, November 2014, pp. 112–119.

[13] C.-C. Lee, C. Busso, S. Lee, and S.S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 1983–1986.

[14] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 183–196, April-June 2013.

[15] L. Bell, J. Gustafson, and M. Heldner, "Prosodic adaptation in human-computer interaction," in *15th International Congress of Phonetic Sciences (ICPhS 03)*, Barcelona, Spain, August 2003, pp. 2453–2456.

[16] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *International conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, May 2004.

[17] M.J. Zaki, "SPADE: An efficient algorithm for mining frequent sequences," *Machine learning*, vol. 42, no. 1-2, pp. 31–60, January 2001.

[18] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C. Wu, and V. S. Tseng, "SPMF: a java open-source pattern mining library," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3389–3393, January 2014.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[20] A. Rosenberg, "AutoBI - a tool for automatic ToBI annotation," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 146–149.

[21] M. W. Meteer, A.A. Taylor, R. MacIntyre, and I. Rukmini, "Dysfluency annotation stylebook for the switchboard corpus," Technical report, Linguistic Data Consortium, Philadelphia, PA, USA, February 1995, Revised June 1995 by Ann Taylor.