TSCR: A comprehensive coding system for task-oriented vocal interaction analysis

Brandon Copping[1], Elizabeth Holzmeyer[1], Santosh Kumar[1], Deniz S. Ones[2], Eugene H. Buder[1]

[1]University of Memphis, Memphis, TN, United States
[2]University of Minnesota, Minneapolis, MN, United States
bcopping@memphis.edu

ABSTRACT

A coding system is developed for the assessment of vocal coordination during task-oriented discourse to systematically assess synchronization and alignment phenomena at multiple levels concurrently. The TSCR implementation includes auditory-spectrographic inspection for Talk unit identifications, Syllable segmentations, Conversational event labels, and Respiratory kinematic signal analysis. TSCR is named from these levels of analysis: talk, syllable, conversation, and respiration. The work-based discourse goals for this corpus required interdependency, and the study incorporated confederates as members of the same-sex dyads who either facilitated or impeded productivity. Participants met each confederate once. Each 30-minute interaction was coded for 3-minute intervals at the beginning, middle, and end. The highest interactivity segments were identified by the product of participants' talk onset frequencies. Alignment results derived from coding of these segments may be interpreted in terms of task efficiency. More broadly, the coding scheme is designed to identify and evaluate physical entrainment dynamics in communicative interactions.

Keywords: Conversation, Turn Exchange, Syllable, Respiration, Entrainment

1. INTRODUCTION

Conversations require every participant to actively listen, predict, and respond to one another in real time. This complicated maneuver requires ongoing coordination between participants, most clearly seen in the turn exchange. The concept of turn-holding can be ambiguous [8] but relatively easy to operationalize dyadically as involving transitions between partners' talk episodes. Numerous past studies have shown that people have remarkable ability to predict when it is they may begin speaking so that they will not overlap too much as to seem rude or interruptive but also not be after too much of a gap as to create an awkward silence [10], [11], [16]. However, how this is performed and what the basis for this synchronization ability is in speech remains a matter of ongoing debate.

There have been many different approaches to analyzing and understanding conversational interaction focusing on turn-exchange dynamics. Sacks et al.'s [18] seminal framework described rules initiated at 'projected transition-relevance places' specifying whom may continue or defer to take a turn, implying physical units defining such places and their projection, but also the units of silence that a turn-holder may use to defer. Similarly, Couper-Kuhlen's [6] groundbreaking discourse analyses distinguished between turn-exchange pauses that were of standard brevity, relative to surrounding speech rhythms, from those that were markedly longer for conspicuous discourse purposes. Meanwhile, Stivers et al. [20] identified turn-exchange pauses cross-culturally as averaging on the order of syllable durations but most typically at 0 ms. This suggests that projection by such units facilitates turn-exchanges, though there is also evidence that other longer prosodic units may be involved [3]–[5] along with respiratory dynamics [12]. Wilson & Wilson's coupled oscillator model for turn exchange [21] provides a contemporary synchronization approach that is useful for modeling such phenomena at multiple levels, indicating that studies might assess whether respiratory versus syllable-level alignments are more significant. For progress to be made on such questions, multiple level descriptions and analyses need to be conducted closely parallel to one another.

The present work introduces a novel, multi-layer coding system for human coding of conversational interactions at three levels, talk, syllable, and conversation, in a corpus that also includes respiratory kinematics. The talk level is the most inclusive, marking any vocal sounds that could have communicative effect and determining the units on which coders at other levels operate. Syllable coders designate the boundaries of all sufficiently articulated vocalizations, accommodating the reductions and hesitations typical of spontaneous interaction to arrive whenever possible at a face-valid count with duration patterning that matches perceived speech rhythms. At the conversational level, which also builds on the scaffolding of talk units, vocalizations are coded for concrete aspects of participants' interactivity. This is done primarily in terms of vocal onsets and offsets relative to the perceived turn-holder, including classification of turn-exchanges as

with or without gaps, or with overlapping between yielder and taker, and noting of simultaneities and interruptive behaviors.

2. METHODS

2.1. Participants and data collection procedure

Recordings included came from 16 sessions: eight same-sex pairs, four male and four female between the ages of 20 and 60, consisting of one participant ("P1") and one or the other trained confederate ("P2"). All were native speakers of American English with a regional, urban dialect reflective of their professional status. Pairs, who had never met, completed a Merit Bonus task: working together on performance evaluations of a set of simulated employee profiles to evaluate and determine an "overall merit score" affecting the employees' year-end salary bonuses [17]. They were told they could accomplish this task in any way they desired and that they would be receiving a compensation bonus depending on how many evaluations were completed across their two sessions combined. These 16 sessions come from a previously recorded corpus of 50 sessions encompassing 25 different participants interacting with the different confederates with these 16 being chosen due to lack of clear recording anomalies or other technical issues with any of their audio or respiratory signals.

In the first session, the facilitating confederate was trained to be *constructive* and helped participants try to complete as many as possible in the 30 minutes allotted. In the second session, their confederate was trained to be *obstructive* and impede productivity. In total there were four confederates, but due to the sessions being sex-matched participants only ever interacted with two of them. Productivity was always higher in the first session, even though being first it required more task orientation. From the point of view of 'complex interdependence,' these circumstances simulate very demanding levels of workplace communication proficiency. After completing the second session, participants were debriefed regarding use of confederates in the design, and they received full compensation regardless of actual productivity in the study.

Conversation partners were seated in a comfortably furnished recording suite with materials on a worktable in between and without any acoustic separation between them. Channels of current interest were acquired by over-the-ear headset microphones (Countryman Associates E6) for full audio, contact microphones adhered to the throat wall for speaker-specific audio (incorporating an accelerometer assembly obtained from PentaxMedical), and respiratory inductance plethysmographic bands [9] for thoracic and abdominal kinesiology (Ambulatory Monitoring Inductotrace systems). Audio from both participants was picked up by the over-the-ear microphones while the contact microphones only picked up signals from the person to whom they were adhered. Audio signals were coded in the AACT coding environment [7], which implements features of the TF32 acoustic analysis program [13].

2.2. Coding procedures

Before any coding was initiated, sessions were split into three segments: the first, middle, and last three minutes of the recording. All segments underwent three rounds of coding. Two of these rounds involved coding only one person at a time, *talk* and *syllable*, while *conversation* coding involved judgments of the interactions within the pairs. *Talk* coding was always performed first as both *syllable* and *conversation* coding used *talk* codes as scaffolding for where work needed to be done. Coding review sessions were held consistently during the process where coders could bring up questions they had during coding to the group and consensus developed as necessary. Statistical reliability assessments were conducted at the conclusion of coding at which point one main coder each remained for the different coding types. Intercoder reliabilities were calculated between these main coders and the same work coded previously by different coders while intracoder reliability was assessed between these remaining main coders and their own previous work. For each type of coding, five one-minute segments were used from the entire corpus to assess reliability.

2.2.1. Talk Coding

The main goal of *talk* coding was to label all sounds within each session partners used communicatively. What could be considered communicative was left purposefully broad to include sounds such as tongue clicks, audible inhales and exhales, and lip smacks, but only if they were deemed by the coder to be intentional, directed, and able to be perceived by the opposing conversational partner. These gestures are being analyzed using the respiratory alignment procedures outlined later in this paper. If such gestures were deemed too minute as to be detected by the other person, they were not coded. Later *syllable* coding was then more restrictive regarding articulated speech.

For segmenting talk units, consistent duration criteria were applied to unarticulated silences: if the silence was less than 200 ms, no gap between codes was allowed and if over 300 ms, a gap between codes was required, but if the silence was between 200 and 300 ms, coders were allowed to consider the salience

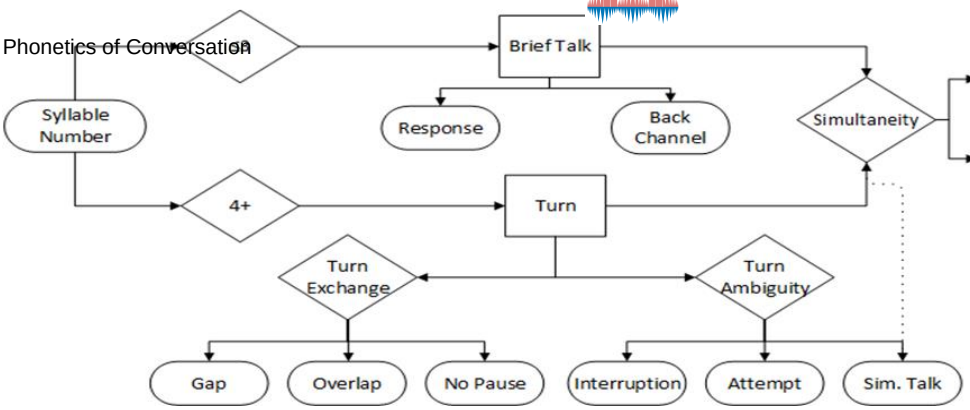

Figure 1: A flowchart of potential *conversation* coding decisions starting with syllable number.

of that gap in their decision, based on such considerations as surrounding speech rate and linguistic discontinuities.

To incorporate laughter, *laugh* and *laughed* codes were available. A talk code could include no more than one syllable of laughter. Ongoing full laughter was coded as *laugh*, and talk that was spoken on laughter was coded as *laughed*; occasionally single talk units were segmented into several such codes.

Reliability was assessed using relative agreement between code onsets with a tolerance of +/- 50 ms. Across the segments used, intercoder reliability ranged between a kappa of κ=0.97 and 0.99 (mean of 0.98) while intracoder ranged between 0.9 and 0.99 (mean of 0.95).

2.2.2. Syllable coding

Similarly to *talk* coding, systematic guidance for *syllable* coding, e.g., lexical or phonological definitions, don't accommodate articulatory rhythms or the variety of spontaneous productions. The criteria applied here were more broadly based on maximal onset and sonority principles, and allowed for extensive reductions, hesitation phenomena (e.g. glottal stopping or fry) or paralinguistic utterances (e.g. laughter, or non-lexical floor-holding 'tunes'): syllable boundaries were coded whenever a distinct rhythmic unit could be both heard and identified spectrographically—the latter was especially important for identifying evidence for reduced articulations, sometimes leading to rejection of a percept that fulfilled lexical expectations but appeared illusory on closer inspection. No specific restrictions were placed on syllable durations; syllables on the order of 50 ms or less occurred, though rarely less (e.g., a syllabic /n/), and drawn-out articulations in the corpus sometimes resulted in syllable durations exceeding 1 s.

Anticipating acoustic modeling of the signal our approach also incorporated "landmark" criteria, for example formant frequency and amplitude inflections and obstruent energy onsets [19]. Training assessments reveal very high agreement rates (over 90%).

Reliability was assessed using relative agreement between code onsets with a tolerance of +/- 50 ms. Across the segments used, intercoder reliability ranged between a kappa of κ=0.92 and 0.99 (mean of 0.95) while intracoder ranged between 0.97 and 0.99 (mean of 0.99).

2.2.3. Conversation coding

Focusing on both individual and dyadic patterns of vocalization, this coding involved the highest number of decisions and used previously coded talk codes as its scaffolding. The primary focus is on turn exchanges and the labeling of brief non-turn utterances, defined as utterances with 3 or fewer heard syllables. Fig. 1 provides a listing of all codes displayed as a decision flow-chart.

Turn exchange codes are based on the timing of the exchange. No code is needed when there is a salient gap between turns, a simple *overlap* code is placed when their turns overlap (and no one has shown interruption), and a *no pause* code is placed when it sounds like there is neither a gap nor an overlap between the two speakers. Signal criteria are consulted, e.g., a tolerance of +/- 300 ms constrains 'no pause' decisions.

Overlapping of speech does not indicate interruption, and to avoid construals of intent, interruption codes were only used when a speaker's turn-holding articulations were cut short along with lexical discontinuity after the partner began speaking concurrently; the speaker who stopped is coded as *interrupted*. If overlapping speech begins concurrently with the partner's onset with a gap having occurred and then is cut short, this is assumed to have been a turn *attempt*. Overlapping speech may also be coded as *simultaneous talking* when a partner joins in during the other's speech, both contributing turns concurrently without either showing articulatory evidence of being stopped by the other.

Utterances with three or fewer syllables spoken by the partner while the other is taking their turn are coded as either *backchannels* or *responses*. A *backchannel* affirms that the partner has the turn, while a *response*, following some question posed by the partner, merely provides information without taking a full turn.

For synchronization study purposes, any audibly *simultaneous onsets* or *offsets* of partners' speech were coded, and two more codes were developed to

reflect the study's task orientation. An *extended silence* was marked whenever interactivity ceased for more than 3 seconds (partners frequently identified independent tasks), and *self-talk* was marked when both manner of speaking and lack of interactivity clearly indicated that the partner was simply 'thinking aloud' during an independent task.

Since conversation coding used talk coding as a scaffold for timing purposes, reliability was instead assessed using agreement on how the interactions were labeled using the various possible codes. Across the segments used, intercoder agreement ranged from 82-96% while intracoder agreement ranged from 87-99%.

2.2.4. Respiratory signals

Thoracic and abdominal wall signals were summed with 2/3 weighting on the thorax [1] and then analyzed in MATLAB for cross-correlations, along the lines pursued by McFarland [12] to assess phase alignments relative to conversational interactions. These cross-correlations were performed using a 10 s sliding window every 5 s.

3. RESULTS

Primary analyses planned for the current corpus of codes and respiratory signals will focus on 25-30 s extracts from each of the three segments of each of the 16 conversations to yield 48 sets of observations, with the goal of comparing partner alignments in constructive versus obstructive sessions. These extracts are selected by a simple algorithm counting the number of turn onsets in each 10 s window, taking the product across participants, and finding the maximally scored interval within each segment. Alignments will be represented primarily in terms of phase-relations between syllable- and respiratory-related units.

To assess the scheme's utility, each of the 48 samples is under examination for interpersonal synchronization phenomena focusing on turn-exchanges, brief talk timing, synchronous behaviors, turn ambiguity, etc. At the most global level, the nature of the turn-exchange flow can be visualized through the respiratory cross-correlation functions. For example, during a task where participants are working simultaneously on a task but not actively exchanging much novel information, respiratory alignment tends to be more in-phase, as can be seen in Fig. 2a This synchrony can rapidly switch to anti-phase on the order of a few seconds when both interlocutors return to exchanging novel information and must therefore alternate with each other, as can be seen in Fig. 2b Since these are conversation samples,

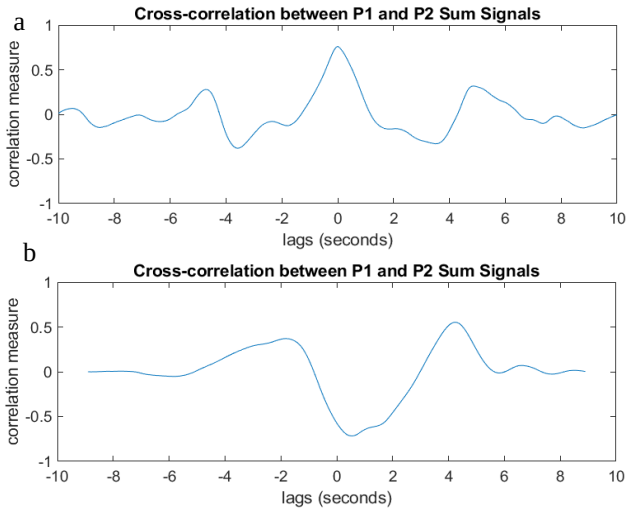

Figure 2a. Respiratory cross-correlation of the first 10 s of a 25 s interactive sample between two female participants showing in-phase alignment.
b. Last 10 s of the same sample as in a, showing that a shift occurred to anti-phase alignment.

this phase switching tends to line up with similar alignments in *talk* codes as well as, during the interactive segments used for analysis, most of their expiratory time is spent speaking. Syllable durations also tend to align during these handoffs with both mono- and di-syllable timing units appearing to be involved. For example, during the final 10 s of the interactive segment referenced in Fig. 2, there is a point where P1 produces 2 ~150 ms syllables simultaneously with 1 ~300 ms syllable from P2, followed by a ~300 ms syllable from P1, and then a similarly-sized syllable from P2 with a "no pause exchange" coded between the final two by conversation coders. This transition was not labeled as an interruption or attempted interruption by any conversation coder and so was judged to be smooth. These multiple levels taken together are what are able to provide more comprehensive insight into conversation dynamics and its relationship to phase dynamics and synchronicity between conversational partners.

4. DISCUSSION

The approach developed here identifies alignments at multiple levels of analysis in task-related discourse. With the present corpus, phase alignments will be examined in relation to task productivity. Furthermore, this framework potentially integrates numerous prior approaches to the phonetics of conversation [15], [22], [23] and accords with coupled-oscillator models [14], [21]. Respiration mobile-sensing models could also be implemented based on these lab data to detect performance-related conversational dynamics remotely [2].

5. ACKNOWLEDGEMENTS

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800006. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.
The authors also thank students and staff at the MD2K Center and the Social Interaction Lab at the University of Memphis for data collection.

6. REFERENCES

[1] R. B. Banzett, S. T. Mahan, D. M. Garner, A. Brughera, and S. H. Loring, “A simple and reliable method to calibrate respiratory magnetometers and Respitrace.”

[2] R. Bari, R. J. Adams, Md. M. Rahman, M. B. Parsons, E. H. Buder, and S. Kumar, “rConverse: Moment by Moment Conversation Detection Using a Mobile Respiration Sensor,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, pp. 1–27, Mar. 2018.

[3] E. H. Buder and J. L. Edrington, “Conversational prosodic interactivity when one partner has aphasia,” in *Speech Prosody 2008 Conference*, P. A. Barbosa, S. Madureira, and C. Reis, Eds. Campinas, Brazil: Editora RG/CNPq, 2008, pp. 501–504.

[4] E. H. Buder and A. Eriksson, “Time-series analysis of conversational prosody for the identification of rhythmic units,” in *Proceedings of the 14th International Congress of Phonetic Sciences*, vol. 2, A. C. Baile, Ed. San Francisco, CA: University of California, Berkley, 1999, pp. 1071–1074.

[5] E. H. Buder and A. Eriksson, “Prosodic cycles and interpersonal synchrony in American English and Swedish,” in *Eurospeech 1997 Proceedings*, vol. 1, E. Dermatas, Ed. Grenoble, France: European Speech Communication Association, 1997, pp. 235–238.

[6] E. Couper-Kuhlen, *English speech rhythm: Form and function in everyday verbal interaction*. Amsterdam: John Benjamins, 1993.

[7] R. E. Delgado, “AACT - Action Analysis Coding and Training Software.” Intelligent Hearing Systems Corp., Miami, FL, 2020.

[8] C. Edelsky, “Who’s got the floor?,” *Lang. Soc.*, vol. 10, pp. 383–421, 1981.

[9] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. al’Absi, and S. Shah, “AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field,” in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, Seattle Washington, 2011, pp. 274–287.

[10] S. Garrod and M. J. Pickering, “The use of content and timing to predict turn transitions,” *Front. Psychol.*, vol. 6, pp. 27–38, 2015.

[11] M. Heldner and J. Edlund, “Pauses, gaps and overlaps in conversations,” *J. Phon.*, vol. 38, pp. 555–568, 2010.

[12] D. H. McFarland, “Respiratory Markers of Conversational Interaction,” *J. Speech Lang. Hear. Res.*, vol. 44, no. 1, pp. 128–143, Feb. 2001.

[13] P. Milenkovic, “TF32.” University of Wisconsin-Madison, Madison, WI, 2018.

[14] M. O’Dell, M. Lennes, and T. Nieminen, “Hierarchical levels of rhythm in conversational speech,” in *Proceedings of the Speech Prosody 2008 Conference*, Campinas, Brazil, 2008, pp. 355–358.

[15] R. Ogden and S. Hawkins, “Entrainment as a basis for co-ordinated actions in speech,” in *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK, 2015, vol. Paper number 0599, pp. 1–5.

[16] J. P. de Ruiter, Holger. Mitterer, and N. J. Enfield, “Projecting the End of a Speaker’s Turn: A Cognitive Cornerstone of Conversation,” *Language*, vol. 82, no. 3, pp. 515–535, 2006.

[17] R. Saavedra, P. C. Barley, and L. V. Dyne, “Complex Interdependence in Task-Performing Groups,” *J. Appl. Psychol.*, vol. 78, no. 1, pp. 61–72, 1993.

[18] H. Sacks, E. A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking for conversation,” *Language*, vol. 50, pp. 696–735, 1974.

[19] K. N. Stevens, “Toward a model for lexical access based on acoustic landmarks and distincitve features,” *J. Acoust. Soc. Am.*, vol. 111, pp. 1872–1891, 2002.

[20] T. Stivers *et al.*, “Universals and cultural variation in turn-taking in conversation,” *Proc. Natl. Acad. Sci.*, vol. 106, no. 26, pp. 10587–10592, Jun. 2009.

[21] M. Wilson and T. P. Wilson, “An oscillator model of the timing of turn-taking,” *Psychon. Bull. Rev.*, vol. 12, no. 6, pp. 957–968, 2005.

[22] M. Włodarczak and M. Heldner, “Respiratory Turn-Taking Cues,” in *Interspeech 2016*, 2016, pp. 1275–1279.

[23] M. Włodarczak, J. Sˇimko, and P. Wagner, “Syllable boundary effect: temporal entrainment in overlapped speech,” in *Proceedings of Speech Prosody 2012*, Shanghai, China, 2012, pp. 611–614.