Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Action-Coordinating Prosody

Nigel G. Ward[1,2], *Saiful Abu*[1]

[1,2]University of Texas at El Paso, [2]Kyoto University
nigelward@acm.org, sabu@miners.utep.edu

Abstract

This paper is an initial exploration of how prosody helps coordinate action, based on examination of speech and motion in a two-player maze game where the players run and jump to avoid obstacles, and coordinate movements to solve problems. We use an unsupervised method, Principal Component Analysis applied to a large set of time-spread features, to discover patterns of behavior involving both prosodic features and game actions. These patterns include prosodic constructions involved in assessing, planning, inhibiting, cuing, and synchronizing actions.

Index Terms: dialog, interaction, joint action, behavior patterns, unsupervised learning, multimodal, missing pitch values, prosodic constructions

1. Motivation

How people manage joint action is an important scientific question [1, 2, 3, 4]. Language often plays an important role; indeed, echoing the old yo-he-ho theory of language origin, Bangerter and Clark have argued that "dialogue has its origins in joint activities, which it serves to coordinate" [5]. The question of how this is done is also of practical interest, for example in human-robot and human-agent interaction [6, 7, 8, 9, 10, 11, 12, 13].

While prosody seems likely to be important in action coordination, to date this has been only peripherally addressed. Work on collaborative tasks such as the Map Task and the Columbia Games Corpus has elucidated, among other things, the prosody of turn-taking [14, 15, 16] and of joint attention [17, 18]. The role of prosody in coordinating turn-taking and feedback has also been studied more generally, not only for its contributions to efficiency but also for its contributions to rapport [19, 20, 21, 22, 23, 24].

Examples of using prosody for control already exist, in systems which directly map user utterance features such as duration, pitch, and loudness to the motion, duration and speed of tools or robots [25, 26]. However such methods have not seen much use, perhaps in part because these mappings are arbitrary and unrelated to how people use phonetics and prosody in interaction with each other [27]. In addition there is work on how virtual agents can align prosody with social and communicative actions, including head nods, eyebrow movements, gaze and gesture [28, 29, 30, 31], but apparently not on aligning prosody with action, either by the agent or by users.

Thus no work to date seems to have directly addressed the question of how people use prosody to coordinate action in the world. This paper accordingly presents an initial exploration of this. The contributions are:

1. A novel corpus of action coordination in gameplay,
2. A demonstration that Principal Component Analysis (PCA) over time-spread features can discover multimodal patterns of behavior,
3. The observation that it is generally easy to identify the pragmatic roles of patterns discovered in this way,
4. The tentative identification of 24 functions served by prosodic patterns in this domain,
5. The finding that prosody does indeed have a role in coordinating action, and
6. A description of 6 prosodic constructions intimately involved in action coordination.

2. Domain

Wanting a domain that involved coordinated action and was easy to instrument, we chose Fireboy and Watergirl, a Flash game free on the Web. This is an easy-to-learn game for two players where each has just three actions: move left, move right, and jump. Nevertheless the gameplay is quite varied, thanks to interesting maze configurations and obstacles, some of which require cooperation to overcome.

We auditioned for someone who enjoyed the game and enjoyed playing it with others, and found one who was alert, positive, helpful and good at keeping it fun. We recorded 19 games of about 10 minutes each, with this one fixed player and 19 partners. We asked them to speak only English, although most were Spanish-English bilinguals. Being more experienced, the fixed player tended to take a guiding role, thus each dialog involves an "expert" and a relative novice. The resulting dialogs turned out suitable for our aims. Compared with, for example, Maptask dialogs, they are fast-paced and the participants are often highly engaged. Samples are available at http://www.cs.utep.edu/nigel/watergirl/. As expected, verbal interaction was important to the game, with the participants making suggestions, discussing the situation, giving warnings, coordinating joint actions, showing excitement, and so on. Also, as anticipated, the participants using highly-varied prosody.

For each game we captured five tracks of information, namely video based on screen captures every 30 milliseconds, head-mounted microphone recordings of each player's voices in separate tracks, and recordings of each player's keystrokes. This data is available for research purposes via the first author.

3. Methods

Our analysis strategy was to compute numerous low-level features at each point in the recordings, and then use PCA to discover the underlying patterns [32, 33].

To represent action we use three features: Running fraction is the fraction of time, in the specified window, during which either the right-arrow or the left-arrow key is held down. Jumps is the count of jumps (up-arrow keypresses) during the window. Motion initiations is the count of keydown events for the right-arrow and left-arrow keys.

For prosody we chose features to cover the four commonly-

 doi: 10.21437/SpeechProsody.2016-129

	number of features		
	expert's behavior	novice's behavior	total
volume	20	12	32
rate	10	10	20
high pitch	14	10	24
low pitch	14	10	24
creakiness	14	10	24
narrow pitch	10	10	20
wide pitch	10	10	20
running fraction	12	12	24
jumps	12	12	24
motion initiations	10	10	20
Total	126	106	232

Table 1: Prosodic and Motion Features

used aspects — intensity, pitch range, pitch height, and speaking rate — plus creakiness. The details of the computation are described elsewhere [34]. One aspect is worth mentioning here, our solution to what Laskowski has called the "missing values" problem [35, 36]. The problem is the general one (becoming more common with the increasing use of machine learning and dimensionality-reduction techniques in the analysis of prosody), that many techniques including PCA, require features which are defined for all inputs, but of course this requirement is not met for pitch in unvoiced regions and regions without speech. Our solution is to not use pitch as a direct feature. Rather, we use pitch information to estimate mid-level features, and then use those in the model. Our mid-level features — pitch highness, pitch lowness, pitch wideness, and pitch narrowness — were chosen as ones that have been implicated in many functions of prosody. We define these features in ways that make them robust to unvoiced regions. For example the value for "high pitch" in a certain window is greater to the extent that the window contains more pitch points that fall in the high band of the speaker's range, and greater to the extent that these points are higher. If there are no such pitch points, because the pitch is either low or non-existent, then there is no evidence for high pitch, and the value is zero. In this way we obtain robust and meaningful values that are defined everywhere, even in unvoiced utterances and regions without speech.

To broadly characterize the pattern of activity in the vicinity of any point in time, t, these features are computed over windows that together tile a span from about 3 seconds before t to 3 seconds after. The window sizes are roughly proportional to the distance from t, thus for example the most distant volume window is 1.6 seconds long, from –3.2 s to –1.6 s, and the closest is 50 milliseconds long, from –50 ms to 0 ms. Windows are fixed in offset from t, rather than being turn-, utterance-, word-, or syllable-aligned, so that they can be everywhere-computable and robust. There are windows for features of both players, enabling the discovery of joint behaviors.

Table 1 lists the numbers of windows for each feature type. There are more for the expert player than for the novice because we wanted to more precisely characterize his prosody in order to eventually build a speech synthesizer with the same expressive abilities. Wanting the top factors to be important ones, rather than those which merely explained the most raw variation, before applying PCA we z-normalized all features. Nevertheless the larger number of pitch windows biases factors which relate heavily to pitch to come out nearer the top, and similarly for factors involving the expert's prosodic behavior. In any case, the exact choice of features to use is not critical to the method. Indeed, in a preliminary study many of the patterns found were largely the same as those described below, despite the use at that time of different pitch-range features, different feature-computing code, different window sizes and offsets, and simplistic substitution for "missing" pitch values.

To discover the patterns, we computed these 232 features at 522,476 timepoints t, sampled evenly every 10 milliseconds through 87 minutes of gameplay. At each timepoint the features characterized the activity in that vicinity, and all this data was fed to PCA to discover the underlying factors.

We then examined the top factors to gain insight into how prosody and action relate in this domain. Each factor has, of course, a loading on the underlying features, and from these it was often possible to directly ascribe a meaning to the factor. For example, Factor 1 loaded positively on all volume features and negatively on all motion features. Thus this factor, on the positive side, describes a pattern in which both players are talking while neither is moving. Conversely the negative side of this factor describes a pattern in which neither player talking while both are moving. Thus this factor reveals two common patterns of interaction, one present in the data at timepoints when this factor has a high value, and the other present when it is negative.

For most factors, however, the loadings are more complex. To understand the corresponding patterns we therefore considered not only the loadings but also what was going on in the game, both at timepoints when a factor had a high positive value and when it had a highly negatively value. Patterns can be present to various extents, but we chose to examine only locations when they were strongly present, thinking that these would be the most informative. For each such extreme-valued timepoint we considered not only the prosody and the keystrokes, but also the state of the game, the words said, their apparent pragmatic intention and effect, and any interesting phonetic aspects. While one can imagine more structured ways to do this, we used a qualitative inductive method, as this was appropriate for an initial exploration of this kind.

4. Prosodic and Action-Prosodic Patterns

This section concisely describes some of the patterns observed. While each is probably best thought of as a complete configuration of features which function together as a whole, for reasons of space, we here describe for each only a few of the the most heavily loaded prosodic and action features. The complete loadings for each factor are at http://www.cs.utep.edu/nigel/watergirl/. For convenience of reference we give in bold a "tagline" for each pattern, referring to some of these features or to commonly co-present pragmatic functions, dialog activities or situational properties. We also mention other commonly co-occuring aspects of the context, phonetics, and words used.

Factor 1 explained 12% of the variance. As suggested above, this involved two patterns or "constructions," one at times when the factor was present positively, and one when present negatively:

1negative **moving and not talking**

1positive **talking and not moving**

Factor 2 explained 8% of the variance. Unusually, it was purely prosodic, with negligible loadings on the action features.

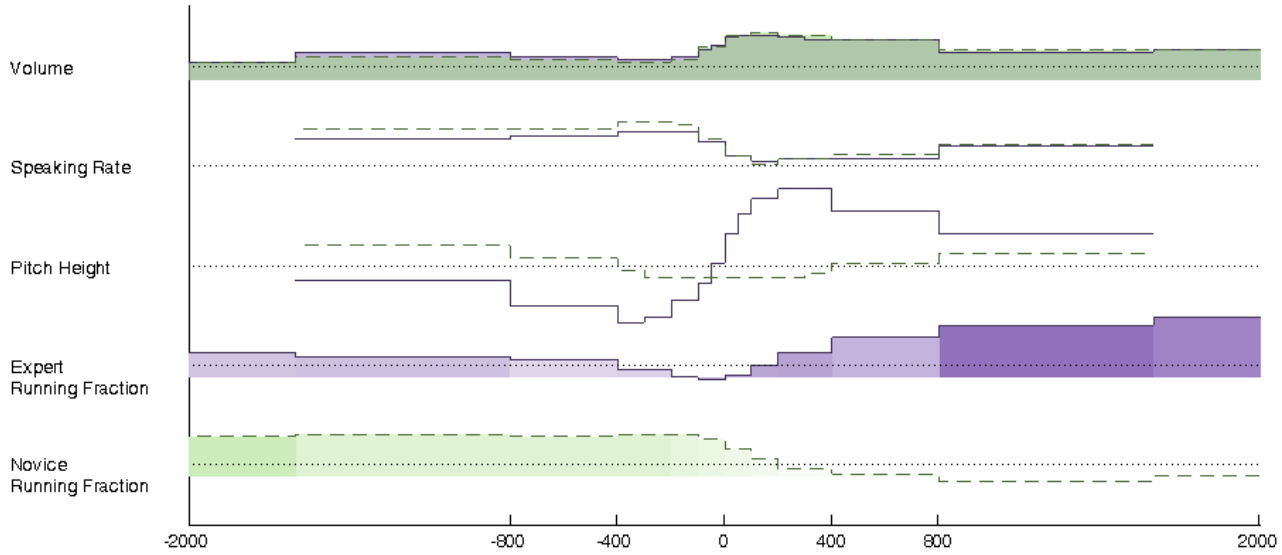

Figure 1: Some loadings of Factor 23. Purple and solid lines are for the expert player; green and dashed for the novice. Time is in milliseconds. The dotted lines are zeros, with points above them positively loaded and points below negatively loaded. The baseline for the shading is -0.1, chosen just to make the patterns visually clearer. The "pitch height" line shows the difference between the loadings of the high-pitch and low-pitch features. The darkness of shading in the running-fraction graphs indicates the loadings on the jumps feature.

2n **predictable, self-directed talk**, by the expert, which was quiet, creaky, and slow.

2p **shared emotional response to an unexpected event**, which was, as always, prosodically the opposite of the negative-side pattern, namely loud, modal, and fast for about a half second

Factor 3 explained 5% of the variance.

3n **pausing to think and plan**, with the expert speaking quietly for a second or two, with a narrow pitch range

3p **moving while talking about something unrelated**, with wide pitch range.

Factor 4 explained 4% of the variance.

4n **concentrating, focusing, each mostly in silence**

4p **moving together wildly while exclaiming excitedly**, with a short burst of high speaking rate

Factor 5 explained 3% of the variance. This also had negligible loadings on the action features.

5n **together making a decision about future action**, with both speakers talking, largely unvoiced or creaky, then slowing in rate for a second, then both falling quiet, speaking quietly in modal voice if at all

5p **digressing to comment on the game or recent performance**, involving silence followed by some creaky, loud commentary, observations, regrets, teasing, etc., with the novice tending to be moving

Factor 6 explained 2% of the variance.

6n **grounding to establish a referent, talking about a future obstacle**, while the expert is moving but the novice is still. Both are talking in quick alternation or with overlap, then the expert's pitch goes low when he identifies the thing the novice was trying to describe.

6p **tense while the novice executes a tricky move** while the expert is still; both very quiet initially, but with a high-pitched comment from the expert, of praise or sympathy, after the novice succeeds or fails

Factor 7 explained 2% of the variance.

7n **fail, apology, minimization**. This pattern is closely related to specific game situations, often occuring when either the novice fails to make a jump and dies, or by inaction causes the expert to die. Prototypically, both players show fear with low pitch, and at the point of failure one or both produce an affect burst [37], in high pitch. Then they both laugh, the novice with an embarrassed laugh and the expert with a sympathetic laugh. They remain motionless waiting for the level to restart, and the novice says something apologetic, using a wide pitch range. The expert then says something brief to play down the need to apologize, then says something louder and more creaky that addresses what to do next and/or provides encouragement. In the data no single sequence exactly matches this prototype, but there are many cases where most of the elements are present.

7p **quiet satisfaction**. This pattern often occurs when the two players have resolved or accomplished something, are pleased with each other, and now just need to do something easy and slow to complete a level. They talk quietly, end in a low-pitched phrase, such as *there you go*, and then fall silent.

Factor 8 explained 2% of the variance.

8n **mock panic**, as for example when the novice is failing to do something, and the expert draws his attention to it with raised pitch and often breathy repetitions, as in *the gem, the gem, the gem* and *ah ah ah, jump jump jump*.

8p **meeting expectation**. This pattern often occurs when the novice is behaving as the expert wants or expects, and the expert marks this with a low-pitch utterance.

Factor 9 explained 2% of the variance.

9n **short instructions**, such as *go for it*, said with no intention of taking the floor, with a decrease in speaking rate and an increase in creakiness.

9p **long instructions**, such as *I need you to push the button so the, so the ledge goes down*, typically with a two-part structure, where there is a creaky region before the end of the first part, possibly signalling the intent to keep the floor.

Most of these factors involve both prosody and action. However their action feature loadings indicate, at best, a moderate elevation or depression of the frequency of some kind of action over some wide region of time. Thus it seems that most of the major prosody-action linkages are just tendencies, individually seldom precise or decisive, although perhaps cumulatively strong. (As these factors are, after all, independent dimensions, any point in time can have loadings on many; and indeed typically several patterns are highly involved at any point in time.)

While an interesting finding, we still wanted to find clearer examples of prosody-action coordination. We continued examining the lower-ranked factors, scanning the loadings of each until we found some with complex temporal configurations of the action features. Coincidentally there were three in sequence: Factors 23, 24, and 25. Each of these, while explaining less than 1% of the variance overall, and only rarely strongly present, accounted for a lot of what was happening at such times.

For Factor 23, Figure 1 shows most of the loadings.

23n **cueing the novice to start moving** by production of a high-pitch phrase by the expert, for example *go for it*.

23p **cueing the novice to stop moving**, where the expert produces a sudden drop in pitch level, followed by the novice stopping motion, and then immediately by a loud high pitch region and a motion initiation by the expert. This sudden drop in pitch level can within a word, for example a stretched out *thennnn* or between words said at a normal rate, for example before the *now* in *okay, so now*. This construction may relate to the common downstep pattern of English [38, 39].

Factor 24

24n **figuring something out**. This pattern occurs when the expert is digesting new information, such as a new level's configuration or a new type of failure, and has figured out what to do next or what just happened. He then typically produces a short phrase like *hmm, okay* and then explains what he figured out or just starts moving.

24p **getting his bearings**. This pattern often occurs when the game situation has just changed, for example after a new level has just loaded, or when the other player has moved into a position enabling the expert to proceed. At times when it is strongly present, the expert often quietly produces thoughtful, filler-type words or phrases like *so, okay, yeah*, and *so, I'm guessing, let's see* with a narrow pitch range. After this the expert doesn't move for a few seconds while he seems to get his bearings.

Factor 25

25n **novice and expert jump in synchrony, novice moving**, often with multiple retries and repeated small adjustments before making a successful jump

25p **expert and novice jump in synchrony, expert running**, typically building up momentum and simultaneously doing a couple of rhythm-establishing leaps before performing one carefully timed jump, with which the novice tends to jump in synchrony. The expert has a pitch dip and rise, followed after about 400ms by a pitch dip and rise by the novice, probably helping synchronize their motion.

5. Summary, Directions for Future Work, and Potential Value

This paper has shown that prosody and action are linked in this domain, shown that PCA can discover multimodal patterns of behavior, and presented descriptions of some prosodic behaviors common in a coordination-rich interaction.

These results are preliminary in several respects: PCA discovers statistical correlations but does not tell us anything about causality; further study is needed to determine whether these prosodic patterns actually have causal efficacy in the coordination of action. Our feature set was designed to capture prosody as it relates to pragmatic and interactional functions; investigation with other features and other methods is needed to determine how the patterns found relate to other prosodic elements and functions.

Future work might also extend this initial exploration in several directions. First, the patterns this method discovers are intrinsically joint, accounting for the coordinated behavior of two speakers, so it would also be interesting to examine these behaviors with traditional approaches, focusing on the decisions and productions of one speaker at a time. Second, this study used PCA as a simple way to decompose the contributions of superimposed factors, but other models might have additional advantages. For example Independent Components Analysis may give a model that is sparser and thus easier to interpret, and soft clustering could give a model with exemplars that are common rather than extreme examples [40]. Third, this study used only prosodic and action features. In this framework it is easy to include more features. It would be interesting to also incorporate lexical behaviors (perhaps using a vector-space model), game-state aspects, gaze, and actions described at a higher level than keystrokes, in order to obtain a more comprehensive view of the behavior patterns. Fourth, it would be interesting to examine the cross-domain generality of some of the pragmatic-prosodic mappings identified, regarding for example grounding, apologizing, and taking the floor.

The potential uses of the patterns discovered here are various. They might be useful for language learners, as such behaviors are not universal [41]. They could be employed by speech synthesizers, to make their outputs prosodically more context-appropriate and effective. They can serve as a list of things that robots and similar systems should to be able to recognize and to produce. They could be used for predicting a player's likely future actions from his or her prosody, or to help a robot player choose his actions based on a human player's utterance prosody. More generally, models of this kind could be useful for creating systems better able to coordinate their actions with human partners[42], for situated-action domains and beyond.

6. Acknowledgements

This work was supported in part by the National Science Foundation as project IIS-1449093 and by a Fulbright Award. We thank Paola Gallardo, Esaul Campos, Raul Alvarez, and Tatsuya Kawahara.

7. References

[1] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: Bodies and minds moving together," *Trends in Cognitive Sciences*, vol. 10, pp. 70–76, 2006.

[2] K. L. Marsh, M. J. Richardson, and R. C. Schmidt, "Social connection through joint action and interpersonal coordination," *Topics in Cognitive Science*, vol. 1, no. 2, pp. 320–339, 2009.

[3] G. J. Stephens, L. J. Silbert, and U. Hasson, "Speaker–listener neural coupling underlies successful communication," *Proceedings of the National Academy of Sciences*, vol. 107, pp. 14 425–14 430, 2010.

[4] M. J. Pickering and S. Garrod, "Self-, other-, and joint monitoring using forward models," *Frontiers in Human Neuroscience*, vol. 8, 2014.

[5] A. Bangerter and H. H. Clark, "Navigating joint projects with dialog," *Cognitive Science*, vol. 27, pp. 195–225, 2003.

[6] A. Nijholt, D. Reidsma, H. van Welbergen, R. op den Akker, and Z. Ruttkay, "Mutually coordinated anticipatory multimodal interaction," in *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Springer, 2008, pp. 70–89.

[7] C. Chao and A. L. Thomaz, "Timing in multimodal turn-taking interactions: Control and analysis using timed petri nets," *Journal of Human-Robot Interaction*, vol. 1, pp. 4–25, 2012.

[8] A. St. Clair and M. Mataric, "Studying verbal feedback in human collaborations to inform robot speech production," in *Int'l Conf. on Collaboration Technologies and Systems (CTS)*, 2014, pp. 135–136.

[9] C.-M. Huang and B. Mutlu, "Learning-based modeling of multimodal behaviors for humanlike robots," in *Human-Robot Interaction*, 2014, pp. 57–64.

[10] H. Cuayahuitl, L. Frommberger, N. Dethlefs, A. Raux, M. Marge, and H. Zender, "Introduction to the special issue on machine learning for multiple modalities in interactive systems and robots," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 4, pp. 1–6, 2014.

[11] G. Bailly, A. Mihoub, C. Wolf, and F. Elisei, "Learning joint multimodal behaviors for face-to-face interaction: performance and properties of statistical models," in *Workshop on Behavior Coordination at HRI '15*, 2015.

[12] Y. Mohammad and T. Nishida, "Learning interaction protocols by mimicking, understanding and reproducing human interactive behavior," *Pattern Recognition Letters*, vol. 66, pp. 62–70, 2015.

[13] J. Hough, I. de Kok, D. Schlangen, and S. Kopp, "Timing and grounding in motor skill coaching interaction: Consequences for the information state," *Proceedings of SemDial 2015 (GODIAL)*, 2015.

[14] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language*, vol. 25, pp. 601–634, 2011.

[15] R. Meena, G. Skantze, and J. Gustafson, "Data-driven models for timing feedback responses in a Map Task dialogue system," *Computer Speech & Language*, vol. 28, pp. 903–922, 2014.

[16] G. Skantze, C. Oertel, and A. Hjalmarsson, "User feedback in human-robot dialogue: Task progression and uncertainty," in *HRI Workshop on Timing in Human-Robot Interaction*, 2014.

[17] G. Skantze, A. Hjalmarsson, and C. Oertel, "Turn-taking, feedback and joint attention in situated human-robot interaction," *Speech Communication*, 2014.

[18] Z. Yu, D. Bohus, and E. Horvitz, "Incremental coordination: Attention-centric speech production in a physically situated conversational agent," in *Sigdial*, 2015.

[19] J. Gratch, N. Wang, A. Okhmatovskaia, F. Lamothe, M. Morales, R. van der Werf, and L.-P. Morency, "Can virtual humans be more engaging than real ones?" *Lecture Notes in Computer Science*, vol. 4552, pp. 286–297, 2007.

[20] N. G. Ward, O. Fuentes, and A. Vega, "Dialog prediction for a general model of turn-taking," in *Interspeech*, 2010.

[21] D. Bohus and E. Horvitz, "Multiparty turn taking in situated dialog: Study, lessons, and directions," in *SIGdial*, 2011.

[22] T. Kawahara, T. Iwatate, and K. Takanashi, "Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations," in *Interspeech*, 2012.

[23] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, Gary, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wollmer, "Building autonomous sensitive artificial listeners," *IEEE Transactions on Affective Computing*, vol. 3, pp. 165–183, 2012.

[24] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency, "SimSensei Kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of AAMAS*, 2014.

[25] T. Igarashi and J. F. Hughes, "Voice as sound: using non-verbal voice input for interactive control," in *Proceedings of the 14th annual ACM symposium on User interface software and technology*, 2001, pp. 155–156.

[26] J. Malkin, X. Li, S. Harada, J. Landay, and J. Bilmes, "The vocal joystick engine v1.0," *Computer Speech & Language*, vol. 25, pp. 535–555, 2011.

[27] L. Vainio, M. Tiainen, K. Tiippana, N. Komeilipoor, and M. Vainio, "Interaction in planning movement direction for articulatory gestures and manual actions," *Experimental Brain Research*, vol. 233, pp. 2951–2959, 2015.

[28] Y. Ding, C. Pelachaud, and T. Artières, "Modeling multimodal behaviors from speech prosody," in *Intelligent Virtual Agents*. Springer, 2013, pp. 217–228.

[29] R. Voigt, R. J. Podesva, and D. Jurafsky, "Speaker movement correlates with prosodic indicators of engagement," in *Speech Prosody*, 2014.

[30] H. Admoni and B. Scassellati, "Data-driven model of nonverbal behavior for socially assistive human-robot interactions," in *16th International Conference on Multimodal Interaction (ICMI)*. ACM, 2014, pp. 196–199.

[31] T. Janssoone, "Temporal association rules for modelling multimodal social signals," in *International Conference on Multimodal Interaction*. ACM, 2015, pp. 575–579.

[32] N. G. Ward and A. Vega, "A bottom-up exploration of the dimensions of dialog state in spoken interaction," in *13th Annual SIGdial Meeting on Discourse and Dialogue*, 2012.

[33] N. G. Ward, "Automatic discovery of simply-composable prosodic elements," in *Speech Prosody*, 2014, pp. 915–919.

[34] ——, "Midlevel prosodic features toolkit," 2015, http://www.cs.utep.edu/nigel/midlevel/, https://github.com/nigelgward/midlevel.

[35] K. Laskowski, "Auto-imputing radial basis functions for neural-network turn-taking models," in *Interspeech*, 2015, pp. 1820–1824.

[36] J. P. van Santen, E. Klabbers, and T. Mishra, "Toward measurement of pitch alignment," *Italian Journal of Linguistics*, vol. 18, pp. 161–187, 2006.

[37] M. Schröder, "Experimental study of affect bursts," *Speech Communication*, vol. 40, pp. 99–116, 2003.

[38] J. Day-O'Connell, "Speech, song, and the minor third: An acoustic study of the stylized interjection," *Music Perception*, vol. 30, pp. 441–462, 2013.

[39] O. Niebuhr, "Stepped intonation contours: A new field of complexity," in *Tackling the complexity of speech*, R. Skarnitzl and O. Niebuhr, Eds. Charles University Press, 2015.

[40] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *European Conference on Computer Vision*. Springer, 2008, pp. 696–709.

[41] N. G. Ward and P. Gallardo, "Non-native differences in prosodic construction use," 2015, submitted.

[42] N. G. Ward and D. DeVault, "Ten challenges in highly-interactive dialog systems," in *AAAI Symposium on Turn-taking and Coordination in Human-Machine Interaction*, 2015.