INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Quantitative analysis of backchannels uttered by an interviewer during neuropsychological tests

Gérard Bailly [1], *Frédéric Elisei* [1], *Alexandra Juphard* [2], *Olivier Moreaud* [2]

[1] GIPSA-Lab, CNRS , Grenoble-INP & Univ. Grenoble-Alpes
[2] CMRR and Neuropsychology, Neurology dpt., CHU de Grenoble
{firstname.familyname}@gipsa-lab.fr, {AJuphard,OMoreaud}@chu-grenoble.fr

Abstract

This paper examines in detail the backchannels uttered by a French professional interviewer during a neuropsychological test of verbal memories. These backchannels are short utterances such as *oui*, *d'accord*, *uhm,* etc. They are mainly produced here to encourage subjects to retrieve a set of words after their controlled encoding. We show that the choice of lexical items, their production rates and their associated prosodic contours are influenced by the subject performance and conditioned by the protocol.

Index Terms: backchannels; neuropsychological test; lexical markers; prosody

1. Introduction

Conversational feedback is most performed though short utterances such as *yeah*, *really*, *okay*, *uhm* produced by interlocutors in order to signal that they are attending the speaker and do not wish to take the floor. Unfortunately these *continuers* [1] can also signal topic shifts as well as trigger propositional content, each function being not mutually exclusive – e.g. *okay* may both signal success and encourage continuation.

Most qualitative studies of backchannels' production have been collected during semi-spontaneous dialogs triggered by conversational themes [2] or games such as map tasks [3] or other collaborative games [1]. Regularities of such multimodal interactive behaviors are mined and – together with data from the literature – often straightforwardly implemented in conversational virtual agents and social robots [4] without considering the very specificity of the task. We analyze here the back-channels produced during short-term face-to-face interviews aiming at evaluating potential deficits of cognitive abilities of interviewees. This study is part of a broader research aiming at giving humanoid robots social skills for monitoring task-oriented interviews. Backchannels are in fact an important factor in creating the impression of cooperative, natural human dialog for synthetic dialog agents.

2. The corpus

The neuropsychological interviews of our corpus were conducted by the third author of this paper [professional neuropsychologist] in the framework of the ANR project SOMBRERO. These interviews are based on the French adaptation [5][6] of the Free and Cued Selective Reminding Test [7] named the RL/RI 16 that uses written words rather than images. It provides a simple and clinically useful verbal memory test for identifying dementia in the elderly. Particularly, it is a sensible tool for the detection of memory dysfunction associated with the early stage of Alzheimer disease – together with other symptoms such as executive and/or instrumental disorders. The RL/RI 16 protocol consists in four phases: (1) the learning of 16 words together with their semantic categories; (2) 3 successive recall tasks (free recall then indexed recall for the unrecovered items) separated with a distractive task (reverse counting); (3) a recognition task involving the 16 items, 32 distractors (16 different words with the same semantic category and 16 true distractors) and (4) a delayed free and indexed recall. The 4th phase was not administrated in the present study. Mnesic performance is evaluated by comparing recall rates of the subject with regards to mean & standard deviations observed within sane control population of the same age interval.

The behavioral data of one unique interviewer – so that adaptive behavior remains consistent across multiple interactions – will serve as demonstration for our humanoid robot (see last section). Therefore, most of our signals are captured from the interviewer's perspective. No invasive sensors are placed on the subjects. The motion of 25 retroflexive markers placed on the plexus, shoulders, head, arms, indexes and thumbs of the professional interviewer were monitored thanks to a Qualysis® system with 4 cameras. A Pertech® head-mounted monocular eyetracker also monitors the gaze of the interviewer (see Figure 1). Speech data are captured via OKMII high-quality ear microphones and recorded synchronously with a side-view video by HD camera.

Figure 1. Visual data. Left: side view from a fixed HD camera. Right: head-related view from the eyetracker scene camera. The dot superimposed to the scene camera features the current gaze fixation point.

Each interview lasts around 20 minutes, comprising the collection of personal records, the core RL/RI protocol and final report of performance. We analyze here almost two hours of multimodal data for the five subjects. Each subject received a 15€ voucher for his/her participation.

Our interviewer uttered 492 backchannels during her 5 interviews. Given the objectives of this joint task, these backchannels mainly fulfill five functions: assessment (61%), incentive (26%), closure of subtask (6%), optional reply (5%) and confirmation (2%).

http://dx.doi.org/10.21437/Interspeech.2016-22

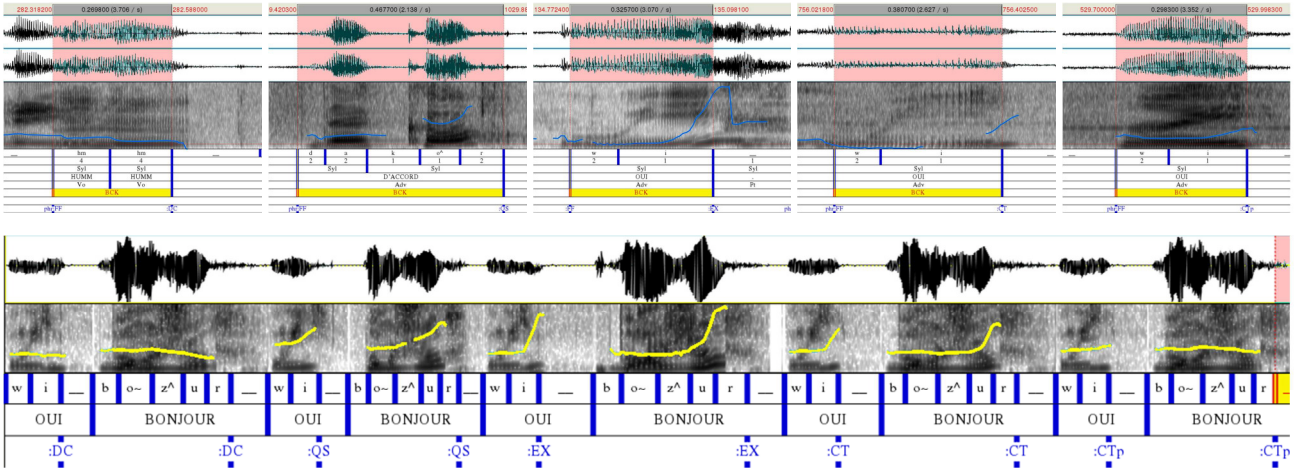

Figure 4. Top. Samples of the 5 types of contours observed in our corpus that are associated with distinctive communicative functions. From left to right: DC (positive assessment with a final F0 fall); QS (full question with a final F0 rise); EX (incentive continuer with a sharp and ample final F0 rise); CT (non incentive continuer with a late final F0 rise); CTp (unmarked backchannel). Bottom: Synthesis of *oui* and *bonjour* using the SFC model trained on our data

We analyze below the lexical markers and the prosodic patterns she used as continuers as well as assessments [8] [9] [10], both being not mutually exclusive.

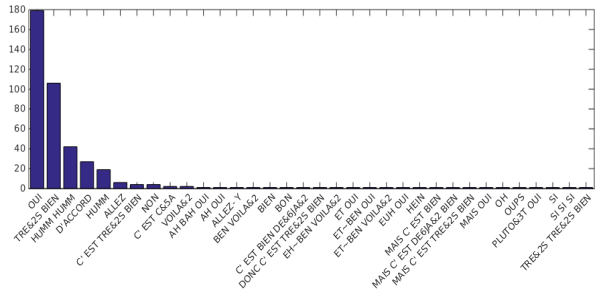

Figure 2. Nb of occurrences of the 34 different lexical markers used by the interviewer to encourage the subjects. This distribution is dominated by 5 items: *oui*, *très bien*, *humm-humm*, *d'accord*, *humm*.

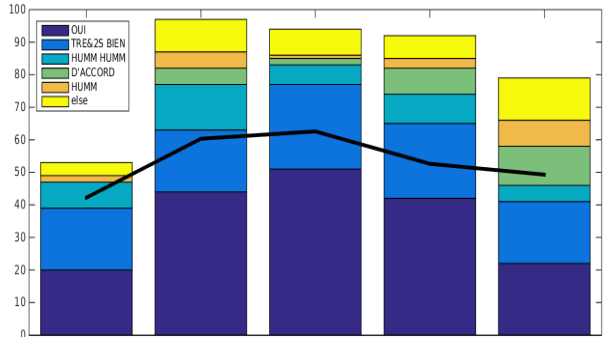

Figure 3. Cumulative histograms of the 5 main lexical markers for each subject. The overlaid black line gives the average number of backchannels per mn (*10). Note that subject 1 has no deficiency of the anterograde verbal memory while subjects 3 and 4 have important mnemonic dysfunctions and deficient semantic encodings. Subject 2 has a deficient semantic encoding. Subject 5 had free recall problems at the 3rd recall task.

3. Lexical markers (LEX)

Prévot et al [11] identified 197 different combinations of basic markers. The most frequent ones were *humm*, *ouais*, *humm-humm*, *ah*, *non*. Our nomenclature (see Figure 2) doesn't differ so much except that positive feedbacks are here boosted: *oui* (yes), *très bien* (excellent), *d'accord* (okay) join up with the basic continuers *humm-humm*, *hum.*

The interlocutor-specific distributions and average number of backchannels per mn (see Figure 3) seems top correlate with subjects' mnesic deficiencies.

4. Prosodic contours (MRK)

In her study of Dutch map-task dialogs, Caspers [12] distinguished between 3 types of melodic contours: H*L L%, LH% and others. Benus et al [1] observed the distribution of boundary tones of backchannels among 4 types of contours: HL%, HH%, LL% and LH%. Backchannels are more likely to have L+H* accents and a high boundary tone (H-H%).

Following the Gestalt approach promoted by the SFC model [13] [14], we labelled our backchannels with 5 distinctive functional markers that both reflect distinctive communicative functions and prosodic patterns (Figure 4):

1. DC encodes a positive assessment with a final F0 fall
2. QS denotes a full question with a final F0 rise
3. EX denotes an incentive continuer with a sharp and ample final F0 rise
4. CT denotes a standard continuer with a final F0 rise
5. CTp cues an unmarked backchannel with a flat F0 contour

Note that the entire corpus was labeled with functional markers that not only include these 5 utterance-level markers but also other utterance-level contours such as QI (that denotes a wh-question usually cued with a final F0 fall) as well as markers related to syntax and emphasis whose contours overlap and add to the utterance-level contours (see [13] for the training framework that disentangles the ill-posed problem of separating overlapping contributions).

Since backchannels are essentially short utterances – often limited to one or two words – they mainly carry one utterance-level marker. The only exception observed in our corpus is

emphasis: 40 adjunctions of broad or narrow focus (notably on the adverb *très*) have been observed (see Figure 5). The distribution of these 5 functional markers over the lexical markers is not arbitrary (see Figure 6): while the incentive continuer dominates because of the interactive task, *d'accord* is the only backchannel used for requesting agreement.

Likelihood ratio tests comparing the combined multinomial model FCT~MRK+LEX with the individual models FCT~MRK and FCT~LEX show that both prosodic (chisq(16)=85, $p<1.e^{-3}$) and lexical (chisq(20)=142, $p<1.e^{-3}$) markers significantly contribute to the encoding of the communicative functions (FCT).

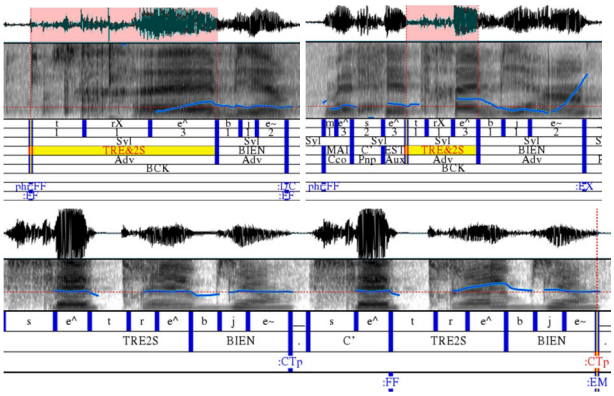

Figure 5. Top: Narrow focus on *très bien* for two utterances in the original data. Bottom: the initial rise on *très* is captured and superimposed on the utterance-level contour (here CTp) by the SFC. Left is the generation of the sole CTp contour. Right is its superposition with EM contour on *très bien*.

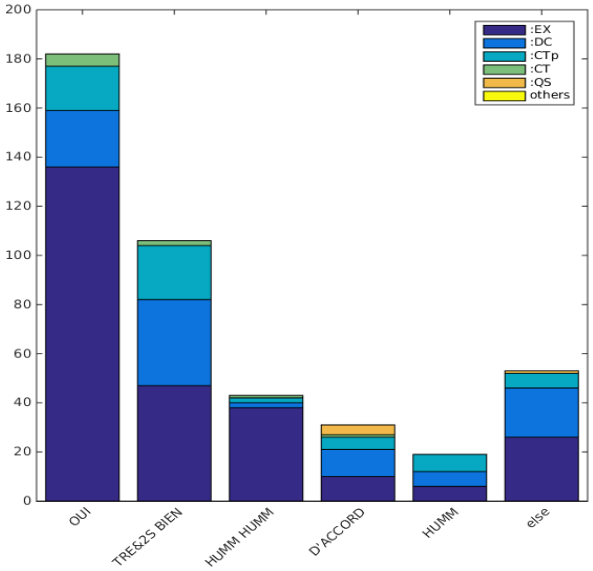

Figure 6. Distribution of the 5 main F0 contours for each main lexical marker. *Humm-humm* is mainly incentive continuer while *d'accord* is sometimes used as a question and *très bien* is equally used as an assessment and a continuer.

5. Phases of the protocol

Figure 7 displays the repartition of lexical markers and prosodic patterns for the different phases of the protocol, i.e. learning, test and recognition (resp. 22.4%, 62.1% and 15.5% of the durations of the interviews on average). The average number of backchannels per mn increases as we progress in the protocol. The use of incentive continuers also strongly increases. The recognition phase generates the highest rate of backchannels per mn, presumably to signal that the 48 subject's decisions have been scored. Note also that the scoring also promotes the use of *d'accord.*

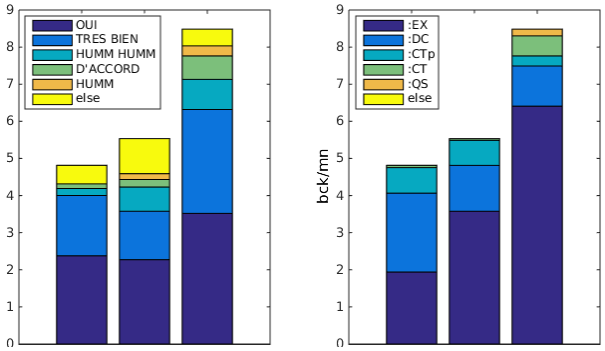

Figure 7. Distribution of backchannels per mn for the three phases (learning, test & recognition) of the protocol according to the lexical markers (left) and prosodic patterns (right).

6. Chaining lexical and prosodic markers

Figure 8 gives the bigrams $p(S_t|S_{t-1}) = p(S_t \& S_{t-1}) / p(S_{t-1})$ for two sets of adjacent tokens S: lexical markers and prosodic patterns. As evidenced in the figure, the choice of lexical markers and prosodic patterns does not only depend on their respective frequency in each phase of the protocol but also obey to syntactic constraints that will be worth replicating by automatic generators of reactive multimodal behaviors.

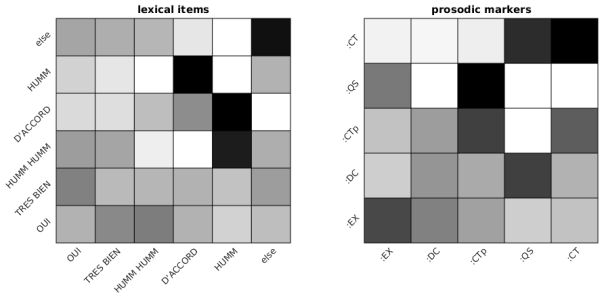

Figure 8. Chaining lexical markers (left) and prosodic patterns (right). We show here the bigrams $p(S_t|S_{t-1})$. S_t are figured in abscissa while S_{t-1} are figured in ordinate. Please note the high conditional probabilities (associated with dark colors) of p(*d'accord*|*humm*), p(*humm*|*d'accord*), p(*humm*|*humm-humm*), p(*else*|*else*), p(CTp|QS) and p(CT|CT).

7. Timing

Figure 9 displays the timing relations between interlocutors' utterances around backchannels. The distributions show that the majority of backchannels are triggered immediately at the end of interlocutors' speeches, presumably to both confirm correct responses and foster further mnesic retrieval. As shown, these backchannels often overlap with interlocutors' speeches. Therefore, an adequate automatic generation of backchannels will need incremental speech recognition technologies in order to react in real-time to relevant spoken input. This contrasts with more informal conversational data that more likely authorizes the modeling of *backchannel*

opportunities (see the review by de Kok [10, pp. 98–101]) using lower-level signal-related cues such as pitch, energy, pause or gaze.

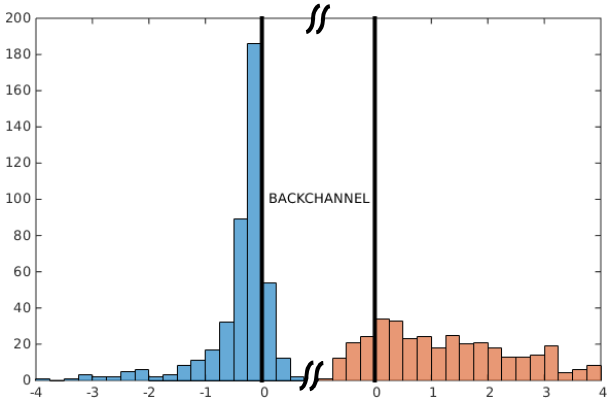

Figure 9. Timing of ends and beginnings of interlocutors' speeches surrounding backchannels. The thick vertical bars align all verbal activities to the onsets and offsets of the backchannels (the distance between them is arbitrary and does not reflect the average duration of backchannels). The majority of backchannels are triggered immediately at the end of interlocutors' speeches.

Table 1. Rates of free vs. indexed recalls, F0 distributions and mean number of backchannels per mn for each subject. Pathological rates are highlighted. F0 is given in cents with reference to 200Hz.

Subject (sex)	Free Recall	Indexed Recall	F0	Interviewer		
				F0	F0 bck	Nb/mn
1 (F)	71	100	1.0 ± 5.3	0.1 ± 4.7	-0.8 ± 6.3	4.22
2 (F)	50	90	-0.7 ± 4.8	1.0 ± 4.9	-0.6 ± 6.2	6.03
3 (M)	40	81	-8.2 ± 4.6	1.6 ± 5.2	-1.6 ± 6.9	6.26
4 (F)	46	100	-1.1 ± 3.6	1.8 ± 4.8	-0.6 ± 6.3	5.26
5 (F)	38	100	-1.0 ± 3.6	1.4 ± 4.6	-1.2 ± 6.3	4.93

8. Inter-subject variability and discussion

Heldner et al [15] have shown that the prosody of backchannels is influenced by so-called backchannel preceding cues (BPC). In fact, backchannels may be privileged locations to observe interlocutors' mutual accommodation, and particularly phonetic alignment. During these interviews, we did not observe any global alignment (see Table 1): the mean F0 of the interviewer has the general tendency to increase as a function of interaction session, except for the last one. This F0 increase as a function of repetitive trials was also observed in several experiments [16] [17]. The mean F0 of feedbacks seems however more correlated to the subjects' registers.

Concerning local correlations, mean F0 of feedbacks moderately correlate (r=0.11) with backchannel following cues, while correlation with BPC does not significantly differ from cues with the same length randomly chosen in the subject's samples. This may reflect the impact of the role of the interviewer who clearly leads the interactions.

We aim at modelling the adaptive communication strategy of one target speaker – who serves as behavioral reference for our conversational agent (see below) – facing multiple interlocutors with various deficits. The data of this one-to-many experimental paradigm are thus clearly idiosyncratic.

9. Conclusions and perspectives

Together with coverbal and non-verbal behavior, backchannels are very important components for humans and agents aiming at fostering physical [18] or mental activity [19] of their conversational partners. We have shown here that neurophysiological tests provide a very interesting framework for studying strategies used by professionals to exploit the various characteristics of these feedbacks, notably rate, placement, prosodic and lexical choice according to their interlocutors and the phase of the protocol. We have shown that these strategies are quite rule-governed and take into account pragmatic needs. One of the main challenge of dialog systems for reproducing these feedback patterns is to be able to incrementally decode and interpret subjects' actions (see notably the proposals made by Schlangen et al [20]).

We have proposed elsewhere a framework [21] [22] to learn coverbal behaviors of one interlocutor (notably gaze and pointing gestures) given the observation of the verbal and coverbal behavior of his/her interlocutor from monitored face-to-face interactions. One of our immediate objectives is to augment the set of output streams with backchannels.

We analyzed here the very first interaction data we collected on face-to-face human-human interviews with a unique professional whose communication skills we have the ambition to endow social robots with. We envision filling the gap between human-human and human-robot interactions with immersive teleoperation (see Figure 10 and [23] [24]) where the communication is mediated by a robotic embodiment that inherently provides perception and action limitations to the higher-level cognitive processes of the pilot. We will now explore the impact of such a robotic embodiment on the behaviors we have characterized here.

Finally these protocol and analysis frameworks are currently being reproduced with several professional interviewers in order to explore the variety of communication strategies. We will notably examine the impact of gender and age on mutual alignment of speech and coverbal behaviors.

Figure 10. Immersive teleoperation of the GIPSA-Lab humanoid robot NINA. The pilot (left) communicates through the sensors (ear microphones and micro cameras embedded in the robot's eyes) and the actuators (neck, eyes, jaw, lips, etc.) of his robotic surrogate (right) with a remote interlocutor.

Appendix

The multimodal data and label files are freely available at:
`www.gipsa-lab.fr/projet/SOMBRERO/data`

Acknowledgements

This research is supported by the ANR (ANR-14-CE27-0014). We thank Sylvain Gerber for the statistical analysis, Ghatsan Hasan for taking care of Nina and our five subjects for their patience and involvement.

References

[1] S. Benus, A. Gravano, and J. B. Hirschberg, "The prosody of backchannels in American English," in *International Congress of Phonetic Sciences (ICPhS)*, Saarbruecken, Germany, 2007, pp. 1065–1068.

[2] P. Blache, R. Bertrand, and G. Ferré, "Creating and exploiting multimodal annotated corpora: the ToMA project," in *Multimodal Corpora. From Models of Natural Interaction to Systems and Applications*, vol. 5509, Michael Kipp, Jean-Claude Martin, Patrizia Paggio and Dirk Heylen, 2009, pp. 38–53.

[3] R. Meena, G. Skantze, and J. Gustafson, "A data-driven model for timing feedback in a map task dialogue system," presented at the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial), Metz, France, 2013, pp. 375–383.

[4] S. Kopp, J. Allwood, K. Grammer, E. Ahlsen, and T. Stocksmeier, "Modeling embodied feedback with virtual humans," in *Modeling communication with robots and virtual humans*, Springer, 2008, pp. 18–37.

[5] M. Van der Linden, F. Coyette, J. Poitrenaud, M. Kalafat, F. Calicis, C. Wyns, and S. Adam, "L'épreuve de rappel libre / rappel indicé à 16 items (RL/RI-16)," in *L'évaluation des troubles de la mémoire : présentation de quatre tests de mémoire épisodique avec leur étalonnage*, M. Van der Linden, Ed. Marseille, France: Solal, 2004, pp. 25–47.

[6] M. Dion, O. Potvin, S. Belleville, G. Ferland, M. Renaud, L. Bherer, S. Joubert, G. T. Vallet, M. Simard, and I. Rouleau, "Normative Data for the Rappel libre/Rappel indicé à 16 items (16-item Free and Cued Recall) in the Elderly Quebec-French Population," *The Clinical Neuropsychologist*, vol. 28, no. sup1, pp. 1–19, 2015.

[7] E. Grober and H. Buschke, "Genuine memory deficits in dementia," *Developmental Neuropsychology*, vol. 3, pp. 13–36, 1987.

[8] C. Goodwin, "Between and within: Alternative sequential treatments of continuers and assessments," *Human studies*, vol. 9, no. 2–3, pp. 205–217, 1986.

[9] J. B. Bavelas, L. Coates, and T. Johnson, "Listeners as co-narrators," *Journal of personality and social psychology*, vol. 79, no. 6, p. 941, 2000.

[10] de Kok, Iwan, "Listening heads," PhD Thesis, U. of Twente, Entschede, the Netherlands, 2013.

[11] L. Prévot, B. Bigi, and R. Bertrand, "A quantitative view of feedback lexical markers in conversational French," presented at the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), Metz, France, 2013, pp. 87–91.

[12] J. Caspers, "Melodic characteristics of backchannels in Dutch Map Task dialogues," presented at the 6th International Conference on Spoken Language Processing (ICSLP), Beijing, China, 2000, pp. 611–614.

[13] G. Bailly and B. Holm, "SFC: a trainable prosodic model," *Speech Communication*, vol. 46, no. 3–4, pp. 348–364, 2005.

[14] B. Holm and G. Bailly, "Learning the hidden structure of intonation: implementing various functions of prosody," in *Speech Prosody*, Aix-en-Provence, France, 2002, pp. 399–402.

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

[16] L. Rantala, P. Lindholm, and E. Vilkman, "F0 change due to voice loading under laboratory and field conditions. A pilot study," *Logopedics Phoniatrics Vocology*, vol. 23, no. 4, pp. 164–168, 1998.

[17] J. A. Jones and K. G. Munhall, "Perceptual calibration of F0 production: Evidence from feedback perturbation," *Journal of the Acoustical Society of America*, vol. 108, pp. 1246–1251, 2000.

[18] J. Fasola and M. Mataric, "A socially assistive robot exercise coach for the elderly," *Journal of Human-Robot Interaction*, vol. 2, no. 2, pp. 3–32, 2013.

[19] J. Le Maitre and M. Chetouani, "Self-talk Discrimination in Human-Robot Interaction Situations for Supporting Social Awareness.," *IJ Social Robotics*, vol. 5, no. 2, pp. 277–289, 2013.

[20] D. Schlangen, T. Baumann, H. Buschmeier, S. Kopp, G. Skantze, and R. Yaghoubzadeh, "Middleware for incremental processing in conversational agents," in *Annual Meeting of the Special Interest Group in Discourse and Dialogue (SIGDIAL)*, pp. 51–54.

[21] A. Mihoub, G. Bailly, and C. Wolf, "Learning multimodal behavioral models for face-to-face social interaction," *Journal on Multimodal User Interfaces (JMUI)*, vol. 9, no. 3, pp. 195–210, 2015.

[22] A. Mihoub, G. Bailly, C. Wolf, and F. Elisei, "Graphical models for social behavior modeling in face-to face interaction," *Pattern Recognition Letters*, vol. 74, pp. 82–89, 2016.

[23] G. Guillermo, P. Carole, F. Elisei, F. Noel, and G. Bailly, "Qualitative assesment of a beaming environment for collaborative professional activities," in *European conference for Virtual Reality and Augmented Reality (EuroVR)*, 2015, p. 8 pages.

[24] G. Bailly, F. Elisei, and M. Sauze, "Beaming the gaze of a humanoid robot," in *Human-Robot Interaction (HRI)*, Portland, OR, 2015, pp. 47–48.