Listen and tell me who the user is talking to: Automatic detection of the interlocutor's type during a conversation

Youssef Hmamouche, Magalie Ochs, Chaminade Thierry, Laurent Prevot

Listen and tell me who the user is talking to: Automatic detection of the interlocutor's type during a conversation

Youssef Hmamouche[1] and Magalie Ochs[2] and Thierry Chaminade[3] and Laurent Prévot[4]

***Abstract*— In the well-known Turing test, humans have to judge whether they write to another human or a chatbot. In this article, we propose a reversed Turing test adapted to live conversations: based on the speech of the human, we have developed a model that automatically detects whether she/he speaks to an artificial agent or a human. We propose in this work a prediction methodology combining a step of specific features extraction from behaviour and a specific deep learning model based on recurrent neural networks. The prediction results show that our approach, and more particularly the considered features, improves significantly the predictions compared to the traditional approach in the field of automatic speech recognition systems, which is based on spectral features, such as Mel-frequency Cepstral Coefficients (MFCCs). Our approach allows evaluating automatically the type of conversational agent, human or artificial agent, solely based on the speech of the human interlocutor. Most importantly, this model provides a novel and very promising approach to weigh the importance of the behaviour cues used to make correctly recognize the nature of the interlocutor, in other words, what aspects of the human behaviour adapts to the nature of its interlocutor.**

I. INTRODUCTION

The *Turing test* introduced by Alan Turing in 1950 aims at testing the conversational abilities of machines. A machine is considered to have passed the test if its answers cannot be distinguished from those of a human being. In this test, the human implied in the conversation is the judge. In this paper, we investigate the idea to analyze directly the human behaviour during a conversation to identify whether she/he talks to an artificial agent or another human.

Recently, several research works on *reverse Turing tests* have been proposed [1], [2] in the research area called Human Interactive Proofs (HIP). The objective is to distinguish human from malicious automated programs. In our research, we aim at proposing another type of "reverse Turing test". Our objective is to give machines the capability to detect if a human is interacting with an artificial agent or a real human. Indeed, our purpose is to provide a model that evaluates automatically the naturalness of an artificial agent conversation capabilities by looking at the user's behaviour. When people interact with such virtual agents, several research works tend to show that users react naturally and socially as they would do with another person [3], [4]. At the same time, one would expect that users' social behaviour, triggered automatically and unconsciously during the interaction [5], are characterized by differences, quantitatively and/or qualitatively, in the social cues such as smiles, head movements or speech activities [6], [7]. However, it is still poorly known *how the social cues expressed by the users interacting with front of an artificial agent differ from those expressed during the same interaction with a real human*, especially in a natural social interaction. Today, we are still far from building artificial agents (robots or virtual characters) capable of participating in a conversation in a *natural* way. Consequently, the human behaviour during a conversation with an *autonomous* artificial agent is different from the one with another human, in particularly, concerning multimodal behavioural cues conveying social information [8], [9]. If autonomous agents were able to converse naturally with users, we could assume that the users' behaviour would be close to the one in human-human conversations. Based on this assumption, we aim at identifying more precisely the behavioural cues that differ between a human-human and a human-machine conversation. For this purpose, the analysis of two multimodal corpora using machine learning methods has been performed. Audiovisual corpora are composed of mediated communication of users interacting either with another human, an embodied conversational agent or a robot talking head in a same context of dialog. The social cues of the users are extracted automatically. These social cues expressed by the users during either a human-human or a human-machine mediated communication are compared using machine learning algorithms. The methodology consists in considering a problem of classification to highlight which social cues differ from one type of interaction to another. Such a methodology has the advantage to enable us to (1) to develop a computational model to automatically detect whether a user is talking to another human or an artificial agent in a mediated communication, by only using information from the user's speech, and (2) to identify the importance of the social cues that distinguishes the two types of interactions.

The paper is organized as follows. In the next section, we introduce our research approach. Section III describes the datasets exploited in this work. In Section, IV, we present the experimental setup and the obtained results. Finally, we discuss the results and conclude in Section V.

[1]International Artificial Intelligence Center of Morocco, Mohammed VI Polytechnic University, Rabat, Morocco. youssef.hmamouche@um6p.ma

[2]Laboratoire d'Informatique et des Systèmes, LIS, UMR7020, Aix-Marseille Université, CNRS,Université de Toulon, 13397 Marseille, France magalie.ochs@univ-amu.fr

[3]Institut de Neurosciences de la Timone, UMR 728. Aix-Marseille Université, CNRS, Université de Toulon, 13397 Marseille, France thierry.chaminade@univ-amu.fr

[4]Laboratoire Parole et Langage, Aix-Marseille Université, CNRS, Université de Toulon, 13397 Marseille, France laurent.prevot@univ-amu.fr

II. OUR PROPOSAL

Mel-Frequency Cepstral Coefficients (MFCCs) are very popular features and very useful in systems that deal with sounds and speeches. They represent a temporal representation of the spectrum of a signal's frequencies. In artificial intelligence field, spectral features such as MFCCs are first generally extracted from input raw signals, then they are used as input of deep learning models (*cf*. Figure 1). MFCCs are computed from more basic feature called filter-banks. MFCCs are just the last step of the conversion process from raw signal into a spectrogram, they can be seen as a compressed form of filter-banks based the discrete cosine transform (DCT). In machine learning, MFCCs are generally used for regression models that require uncorrelated input variables, but this is generally not required for deep learning models. In contrary, it is better sometimes to use the original filter-banks. There are other deep learning approaches based on end-to-end models, which work on raw signals directly without using any predefined features. For example, some architectures based on 3D Convolutional Neural Networks (CNNs) and Convolutional LSTM (Convolutional Long Short Term Memory) are used with the idea of extracting spatio-temporal features within the model itself [10], [11]. But until now, the first approach, which is based on extracting spectral features like filter banks and MFCCs, represents the state-of-the-art approach, and it is the most used in applications related to speech and acoustic signals, such as acoustic signatures classification, speaker recognition, and automatic speech recognition.

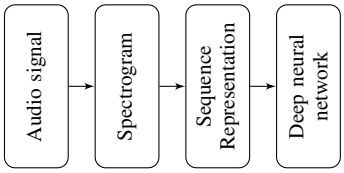

Fig. 1. An illustration of the traditional methodology when performing prediction tasks using speech data.

The disadvantage of using MFCCs or filter-banks is their limited *explainability*, *i.e.*, even if they may allow to have good prediction's accuracy, the results will be difficult to interpret. In other words, it will be difficult to identify explainable causal relations between the speech behaviour (characterized by the extracted speech features, input of the prediction model) and the type of the interlocutor (output of the prediction model). In this work, we adopt a different approach. We are interested in two criteria: (i) Making precise predictions, and (ii) identifying the most relevant linguistic and semantic features that potentially differ between human-human and human-machine conversations. Therefore, we designed new *high-level features* that can be computed from transcriptions of raw speech. They represent highly relevant features of a conversation based on a linguistic research. They are used as intermediate features within a deep neural network as illustrated in Figure 2.

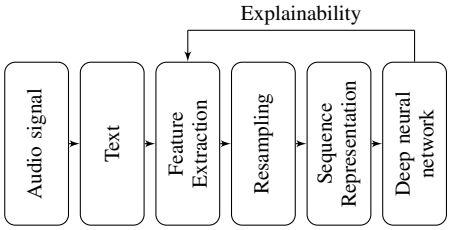

Fig. 2. An illustration of the proposed methodology. First, the speech is transformed to text. Then, specific features are extracted from the obtained text. Finally, the obtained features are re-sampled and reorganized in a form of sequences and put as input of a recurrent neural network.

A. The feature extraction step

The first step is transforming the input speech signal into text. Then, automatic annotation and segmentation are applied using SPPAS [12]. From the obtained transcriptions, *twelve linguistic features* have been defined and extracted as time series. They consist of the *Speech activity* (the presence of speech), *Speech rate* (the speed of the spoken words). Lexical complexity is considered based on two measures, the first is based on the amount of spoken tokens divided by the total number of tokens (*Type-token-ratio*, and the second one (*Lexical richness*) is computed considering the amount of spoken adjectives and adverbs divided by a total number of extracted tokens (including adjectives, adverbs as well as auxiliary words, conjunction, determiners, nouns, prepositions, pronouns, verbs) [13]. We also considered *Filled-pauses* (*i.e.*, utterances like "mmh" during pauses of active speech) [14], and lexical *Feedback* items, representing expressions to communicate perception and understanding, as well as reactions to what the conversational partner had said (*e.g.*, "yes", "no", "okay", *etc*.) [15]. We also included *Discourse markers*, which are expressions used to make the discourse organized (*e.g.*, I mean, so, therefore, okay) [16], and *Spoken particles items* (*e.g.*, but, well, maybe), and also *Laughter*. Sentiment analysis is taken into account based on two metrics; *Subjectivity* and *Polarity* of speech with the Pattern library [17]. These features detect positive and negative feelings and opinions from speech. Polarity is related to positive and negative feelings or opinions, such as anger (negative feeling) or happiness (positive feeling). Values range from 1 (expression of positive feeling) to -1 (expression of negative feeling). Subjectivity is related to subjective content with scores between 0, for objective content, and 1 for subjective content. Table I summarizes all the extracted features and their description.

B. The prediction step

The prediction task can be formulated as a supervised learning problem aimed at associating each sequence of user's features to a binary variable representing the interlocutor's type (human or an artificial agent). Let y_t be the target binary variable. It represents the interlocutor's type (1 for human and 0 for artificial agent). And let $x(t) = \{x_1(t), x_2(t), \ldots, x_k(t)\}$ be the set of k predictive variables representing the input features, where $x_i(t)$ is the i^{th} variable

Features	Descriptions	Details
Speech Activity	The interlocutor speaking?	Based on time-aligned IPU. An IPU is a speech block bounded by pauses and coming from a single speaker [18].
Laughter	Laughter occurrences	Based on word-level time-aligned transcripts
Filled-pauses	Filled-Pauses occurrences	Based on word-level time-aligned transcripts : *"euh", "heu", "hum", "mh"*
Feedback	Conversational Feedback occurrences	Based on word-level time-aligned transcripts : *'oui' (yes), 'ouais' (yeah, 'non'(no), 'ah', 'd'accord'(right), 'ok'* + Laughters
Discourse-Markers	Occurrence of words used to keep speech organized	Based on word-level time-aligned transcripts : *'alors'(so), 'mais'(but), 'donc'(therefore), et('and'), 'puis'(then), 'enfin'(finally), 'parceque'(because), 'ensuite'(after)*
Spoken-Particles	Occurrence of (final) spoken particle items	Based on word-level time-aligned transcripts : 'quoi', 'hein','ben','bon'(well), mais (but), 'bref' (in short)]
Interpersonal	Merge of inter-personal linguistic features	Merge of (Filled-pauses, Feedback, Discourse Markers, Spoken Particles and Laughter)
Speech rate	The speed of the spoken words	Based on time-aligned IPU transcript.
Type-Token-Ratio	Lexical richness measure	Based on time-aligned transcript: (number of different tokens) / (total number of tokens).
Lexical-Richness	Lexical richness measure [13].	Based on time-aligned transcript: (number of adjectives + number of adverbs) / (total number of tokens).
Polarity & Subjectivity	Sentiment analysis metrics [17].	Based on time-aligned transcript, and a pre-trained KNN classifier.

TABLE I

THE EXTRACTED FEATURES.

at time t. We aim at predicting y at time t based on a sequence of x of length p, where p is the lag, or the look-back parameter. This parameter depends on the time-stamp of the features and the chosen sliding window lentgh. In our case, the features are re-sampled with a time-stamp of 0.5s, and the sliding window length is $40s$. Therefore, p is equal to 80 (see more detailed description in the experience section). The following notations are considered to represent the sequences of the predictive variables in a compact way:

$$x_i^{t-p:t} = [x_i(t-p), x_i(t-p+1), \ldots, x_i(t)]^\top. \quad (1)$$

Therefore, our prediction model can be written as follows:

$$y(t) = f\left(x_1^{t:t-p}, \ldots, x_k^{t:t-p}\right) + e_t, \quad (2)$$

where f is function of the model, and e_t is its error vector.

RNNs are particularly adapted to this kind of sequence prediction, since they take into account the temporal aspect of the sequences, and more essentially, they have have a notion of hidden state, which help them to have a sort of memory, and this is their main specificity over other artificial neural networks. We focus in this work on the LSTM, which is one of the main recurrent neural networks used for learning from long sequences, thanks to its forgetting mechanism. It has shown good results in many applications such as time series prediction and neural translation (sequence-to-sequence models).

III. DATASETS

The corpus is composed of data that was recorded as part of an experiment that compares behavioural and physiological responses when a participant has a natural social interaction with a human or an artificial agent (a virtual agent or a robot) in French.

Note that we have created our own database to have *parallel corpora* : conversations of users with either a virtual agent, a robot or another human in a comparable situations. Experimental procedures used for data collection adhered to the Declaration of Helsinki. Note that participants signed informed consent.

Participants are made to believe they participate in a marketing experiment to validate an incoming advertising campaign. The cover story provides a common goal for the two interacting agents as well as a topic for the discussion. A video-conference set-up is used. Each conversation lasts 1 minute. The participants don't know each other.

Corpus 1: human-human and human-virtual agent conversations. The embodied conversational agent used to interact with the participants was the Greta/VIB system [19]. Greta/VIB is an experimental platform specifically dedicated to investigate verbal and nonverbal aspects of human-machine interactions and is particularly relevant for the current project as it is able to reproduce human emotional states and generic behavioural feedbacks [20]. A voice synthesizer from company CereProc was used to generate speech [21]. The embodied conversational agent, presented as autonomous, is used to compare the behavioural responses to interactions with a fellow human. A simple Wizard of Oz (WoZ) procedure controls the ECA, so that unknown to the participant, a human controls the ECA directly. Eleven participants (female students, mean age 22.7 years, SD 6.4 years). interact with another participant or the ECA three times. In total, this corpus is composed of 32 human-human mediated conversations and 32 human-virtual agent mediated conversations of 1 minute.

Corpus 2: human-human and human-robot conversations: In this corpus, we used the FURHAT conversational robot [22]. Furhat was controlled by the confederate in a *Wizard-of-Oz* mode, unknown to the participants who believed it was autonomous. 24 participants underwent each 24 1-min conversations of 1 min, 12 with a human and

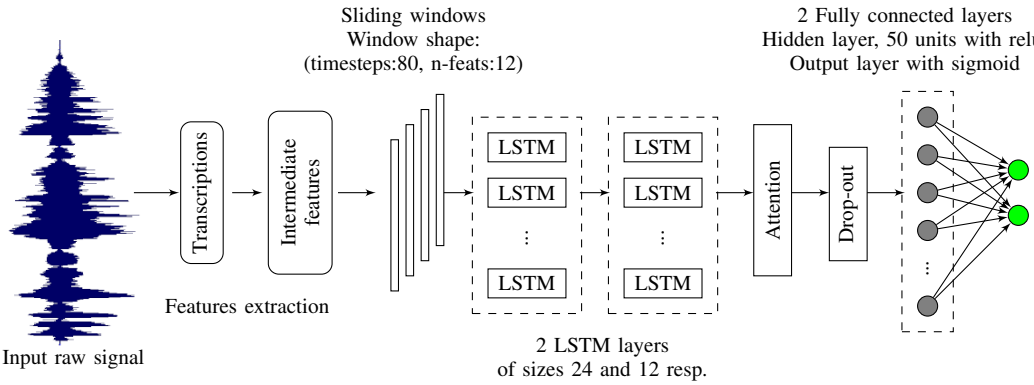

Fig. 3. The proposed artificial neural network architecture. The input consist of sequences of the 12 features. The two hidden layers contain 24 and 12 LSTM blocks respectively, followed by a multiplicative attention layer, and a drop-out layer with a ratio of 20%

as many with the conversational robot, in alternation. The users interacted with the human or robot talking head in a video-conference setup, as for the virtual agent experiment. In contrast to the first experiment though, the user was lying supine in an fMRI scanner while his brain activity was being recorded during the interactions. Another difference is that the interlocutor was an experimenter, unknown to the participant who believed she/he was another naive participant (T) or a confederate of the experimenter for female users. In total, this corpus is composed of 72 human-human mediated conversations and 72 human-robot agent mediated conversations of 1 min. The corpus is described in more details in [23] and the data public availability is available at https://hdl.handle.net/11403/convers/v2.

IV. Experiments and Results

The experiments are performed on the two corpora presented in the previous section. Given the different sizes of the two corpora, we decided to use the corpus with the robot talking head (corpus 2) for the training and validation and the corpus with the virtual agent (Corpus 1) only for the test. The performance of the predictions are presented in the next section. The training and validation sets are both balanced, *i.e.*, the numbers of human-human and human-robot/virtual agent conversations are equal. The training, validation and test data sets are defined in order *to never mix* the two corpora. This way, we make sure that the training and test sets are different in terms of participants, this will help to avoid the overfitting problem, and we can suppose that the obtained results doesn't depend on the experimental set-up.

Feature extraction: We extracted features from raw audios as time series for each participant, based on types of features, the features proposed in our approach (*cf.* Section II), and the filter-bank features, in order to compare both approaches. Note that the proposed model is language independent. In this study, we use the French language, but the implementation also allows to extract features from English audio recordings. Other languages can easily be added to the provided implementation by representing the features in Table I according to the desired language.

Prediction set-up: Based on the model presented in II-B, we choose a sliding window of size 40s to generate the sequences, and since the used time-step to re-sample the time series is 0.5 seconds, the lag parameter, *i.e.*, the size of the sequences, is 80. We think that considering short time lags in this task is not suitable since we have to consider long enough time to be able to see variations in speech features. In the same time, we can't consider much higher lags since the data contain audios file with a duration of one min. Finally, the training and test sets contain 46201 and 6012 observations respectively.

Evaluations: In this part, we compare the proposed architecture (*cf.* Figure 3) with the classical one, which is based on spectral features as discussed in Section II. Within each architecture, we evaluated different types of classifiers, including the Random Forest classifier as a baseline, and three artificial neural networks:

- *A Feed-forward Neural Network (FNN)*: this is the most trivial approach that works by flattening all temporal variables and considering them as the input to a fully connected neural network;
- *A convolutional neural network (CNN)*: this should be an improvement from the previous models since we are dealing with long sequences. The idea is to consider the input sequences as an image where the number of rows is the number of features and the number of columns is the length of the sequence (80). It is composed with a 2 convolutional layers with 32 filters, each followed by a max-pooling layer, then a drop-out layer, and finally fully connected network for classification. See link in note 1 for more details;
- *An LSTM and multiplicative attention based network (LSTM-Attention)*: an architecture based on LSTM layers described illustrated in Figure 3. The multiplicative attention layer ([24]) is included to help the network to focus on the most important parts of the sequences.

Model	Accuracy	Recall	Fscore
LSTM	**0.78**	**0.78**	**0.78**
LSTM-Attention	0.77	0.77	0.77
CNN	0.75	0.75	0.74
FNN	0.74	0.74	0.74
RF	0.65	0.66	0.65

TABLE II

PREDICTION SCORES - LINGUISTIC FEATURES

Model	Accuracy	Recall	Fscore
LSTM	0.62	0.62	0.62
LSTM-Attention	**0.65**	**0.61**	**0.66**
CNN	0.31	0.47	0.49
FNN	0.57	0.57	0.56
RF	0.44	0.50	0.51

TABLE III

PREDICTION SCORES - FILTER-BANK FEATURES

- *An LSTM based network (LSTM)*: an architecture based on LSTM layers as illustrated in Figure 3 without the attention layer.

The neural networks models are implemented based on Tensorflow library version 2.9.0. The networks are trained over 20 epochs, the batch size used is 32, and the stochastic gradient descent algorithm is adopted with a learning rate of 0.01. Codes and data are available in this Github repository[1].

Performance scores on test data: For the Random Forest classifier, a 10-fold-cross validation is performed for parameter tuning. For neural network models, this procedure will take a lot of time, so we did a simple cross-validation with fixed architectures. Tables II and III contain the performances of the evaluated models with the proposed features and the filter-bank features respectively. As commonly used, we have computed three measures to evaluate the quality of the model : the accuracy (or classification accuracy) represents the number of correct predictions from all predictions made; the recall a measure of classifiers completeness, and the Fscore the balance between the precision and the recall. First of all, the results show that the learned models based on LSTM network can predict accurately, compared to the evaluated models, the type of interlocutor (real or artificial) of the users based only on the users' speech behaviour, both based on linguistic features and on filter-bank extractions. The results also demonstrate that the multiplicative attention layer improves slightly the accuracy only with filter-bank features. Even if the training and validation data-sets do not include conversations with the same artificial agent, the model can determine if the user is speaking to a human or an artificial agent. Indeed, in our experiment, the training and validation sets contain conversation with artificial agent that correspond to a robot talking head and the test set contains conversation with a on-screen virtual agent. These results tends to show **the existence of specific users' specific speech behavioural cues during unnatural conversations with artificial agents**. Moreover, the results show clearly that all models performs better when using the linguistic instead of the filter-bank features. These results comparing low versus high level features strongly favor the extraction of known reliable features to improve the performances of the prediction but also the explainability of the model.

[1]https://github.com/Hmamouche/HumanOrRobot

Features	Importance scores
Polarity	0.15
Spoken-particles	0.14
Type-Token Ratio	0.11
Lexical richness	0.10
Interpersonal	0.10
Subjectivity	0.10
Filled-pauses	0.08
Discourses markers	0.07
Speech Activity	0.06
Feedback	0.04
Speech rate	0.02
Laughs	4.54e-05

TABLE IV

THE OBTAINED FEATURES IMPORTANCE.

Features importance: The importance scores of the input features providing the best prediction scores has been computed to better understand the causal relations between the speech features and the nature of the of interlocutor. These scores are computed by normalizing the weights of the first LSTM layer of the network. The obtained scores are presented in Table IV. The interesting aspect in these results is that the most relevant features are not strongly related to the quantity and the speed of speech (speech activity and speech rate features respectively), but to social and linguistic elements of the interaction, *e.g.*, polarity, spoken particles, lexical richness, interpersonal, and subjectivity. These results may be explained by the limited social competencies of the artificial agents used in these two experiments, despite the wizard of oz set-up allowing for correct interactions in terms of their content, and thus the importance of the social dimension in conversations. Consequently, the differences on this social aspect reflected through the users' speech behaviour (polarity, subjectivity and interpersonal features) should be particularly prominent when comparing directly the human-human and human-machine conversations on one subject and in one given context. Concerning the importance of the linguistic features (spoken particles, lexical richness), we can suppose that a human has a more constructed and rich dialog with another human than with the artificial agents, a phenomena explained here again by the poor conversational capabilities of the artificial agents used in these experiments, and reflected in our study through the particularly discriminating weight of theses linguistics characteristics.

V. CONCLUSION

In this paper, we investigated a new application of behavioural analysis during human-human and human-machine interactions: the automatic prediction of the interlocutor's nature (human or artificial) during a mediated conversations solely based on features extracted from the users' speech. In order to develop prediction model that can be interpreted in terms of the artificial agent speech abilities, we have

identified a set of high-level features reflecting the social and linguistic richness of conversations. First, the performances of the learned models show that, in fact, we can accurately predict the nature of the interlocutor based on the user's speech features using a LSTM network. Secondly, the proposed informed features extraction outperform the classical approach based on filter-bank features in terms of prediction performances. An analysis of the features importance moreover identifies differences in specific aspects of the users' speech depending on the interlocutor's human or artificial nature. The user depicts different speech behaviours during a conversation with an artificial agent compared to a human fellow, in particular when the polarity, subjectivity and richness of her/his vocabulary are taken into account.

If we consider the human-human interaction as the natural, or social, one, our research results mean that the artificial agents (virtual agent or robot talking head used in our experiments) are not able to generate natural and social behaviour on behalf of the human user as another humans can do. This result is not totally surprising since these artificial agents are far from comparable to human in terms of social and conversational and communicative, or social, capabilities. However, the proposed model may also be used to evaluate an artificial agent: we can suppose that the more difficult its will be to use such type of machine learning model to distinguish between human-human and a human-machine interactions, the more the artificial agent will be successful in its capacity to trigger social behaviours.

In other words, in this paper, we have explored the possibility to develop an automatic metric for a "reverse Turing test" by solely looking at users' produced speech. The presented research has several limits, in particular given the specific context of communication (conversational contexts, number and nature of the the participants in the experiment, appearances and behaviour of the artificial agents, *etc*., ...). Others experiments with different contexts and artificial agents depicting improved social behaviours need to be conducted to validate and generalize the findings. Moreover, other features considering the multi-modal aspects of the communication (as for instance the user's head movements, that are also sufficient to infer the nature of the interlocutor as demonstrated in another analysis of the first corpus [9], focusing on facial expression, but also the gaze or the physiological responses that should also be explored with the the existing corpora) should also further be explored to fully characterize requirements for improving the social competence of interacting artificial agents.

ACKNOWLEDGMENT

This research is supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and ANR-11-IDEX-0001-02 (A*MIDEX).

REFERENCES

[1] G. Kochanski, D. Lopresti, and C. Shih, "A reverse turing test using speech," 2002.

[2] Y. Rui and Z. Liu, "Artifacial: Automated reverse turing test using facial features," *Multimedia Systems*, vol. 9, no. 6, pp. 493–502, 2004.

[3] B. Reeves and C. Nass, *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press Cambridge, UK, 1996.

[4] N. C. Krämer, "Social effects of virtual assistants. a review of empirical results with regard to communication," in *Proceedings of the international conference on Intelligent Virtual Agents (IVA)*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 507–508.

[5] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *Journal of social issues*, vol. 56, no. 1, pp. 81–103, 2000.

[6] N. Krämer, S. Kopp, C. Becker-Asano, and N. Sommer, "Smile and the world will smile with you—the effects of a virtual agent's smile on users' evaluation and behavior," *International Journal of Human-Computer Studies*, vol. 71, no. 3, pp. 335–349, 2013.

[7] J. N. Bailenson and N. Yee, "Digital chameleons automatic assimilation of nonverbal gestures in immersive virtual environments," *Psychological science*, vol. 16, no. 10, pp. 814–819, 2005.

[8] S. Oviatt and B. Adams, "Designing and evaluating conversational interfaces with animated characters," *Embodied conversational agents*, pp. 319–343, 2000.

[9] M. Ochs, N. Libermann, A. Boidin, and T. Chaminade, "Do you speak to a human or a virtual agent? automatic analysis of user's social cues during mediated communication," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 197–205.

[10] D. Palaz, M. M. Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4295–4299.

[11] T. Parcollet, M. Morchid, and G. Linares, "E2e-sincnet: Toward fully end-to-end speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7714–7718.

[12] B. Bigi, "Sppas-multi-lingual approaches to the automatic annotation of speech," *The Phonetician*, vol. 111, no. 112, pp. 54–69, 2015.

[13] M. Ochs, S. Jain, and P. Blache, "Toward an automatic prediction of the sense of presence in virtual reality environment," in *Proceedings of the 6th International Conference on Human-Agent Interaction*. ACM, 2018, pp. 161–166.

[14] M. Swerts, "Filled pauses as markers of discourse structure," *Journal of pragmatics*, vol. 30, no. 4, pp. 485–496, 1998.

[15] A. Gravano, J. Hirschberg, and Š. Beňuš, "Affirmative cue words in task-oriented dialogue," *Computational Linguistics*, vol. 38, no. 1, pp. 1–39, 2011.

[16] D. Schiffrin, *Discourse markers*. Cambridge University Press, 1987, no. 5.

[17] T. De Smedt and W. Daelemans, "Pattern for python," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2063–2067, 2012.

[18] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[19] C. Pelachaud, "Studies on gesture expressivity for a virtual agent," *Speech Communication*, vol. 51, no. 7, pp. 630–639, 2009.

[20] M. Ochs, R. Niewiadmoski, and C. Pelachaud, "How a virtual agent should smile? morphological and dynamic characteristics of virtual agent's smiles," in *Proceedings of the international conference on Intelligent Virtual Agents (IVA)*. Springer Berlin Heidelberg, 2010, pp. 427–440.

[21] M. P. Aylett and C. J. Pidcock, "The cerevoice characterful speech synthesiser sdk," in *IVA*, 2007, pp. 413–414.

[22] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive Behavioural Systems*, ser. Lecture Notes in Computer Science, A. e. a. Esposito, Ed. Springer Berlin Heidelberg, 2012, pp. 114–130.

[23] Rauchbauer Birgit, Nazarian Bruno, Bourhis Morgane, Ochs Magalie, Prévot Laurent, and Chaminade Thierry, "Brain activity during reciprocal social interaction investigated using conversational robots as control condition," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 374, no. 1771, p. 20180033, Apr. 2019.

[24] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.