# Probing the Inductive Biases of a Gaze Model for Multi-party Interaction

Léa Haefflinger, Frédéric Elisei, Brice Varini, Gérard Bailly

# Probing the Inductive Biases of a Gaze Model for Multi-party Interaction

Léa Haefflinger
GIPSA-lab, Univ. Grenoble Alpes, CNRS, Grenoble INP
Grenoble, France
Atos
Échirolles, France

Frédéric Elisei
GIPSA-lab, Univ. Grenoble Alpes, CNRS, Grenoble INP
Grenoble, France

Brice Varini
Atos
Échirolles, France

Gérard Bailly
GIPSA-lab, Univ. Grenoble Alpes, CNRS, Grenoble INP
Grenoble, France

## ABSTRACT

The behavior management controls proposed for social robots are mostly designed for highly controlled scenarios. In the real world though, robots have to adapt to new situations, generalizing learned behaviors. To address this adaptation challenge, neural network models with embedding layers could be used. We present here an approach to better understand the inductive biases of our robotic gaze model. It was trained with multimodal features as inputs – either endogenous or exogenous to the robot. Inductive biases were explored by observing feature representations in the embedding spaces. We found that the model was able to distinguish between the robot speech intentions that either request or provide information. Similarly, pairs of partners seem grouped according to their social behavior (speaking time, gaze). Finally, we checked that these groupings had a real impact on the model's performance. Driving these biases when facing new people should allow to generate adapted behavior.

## CCS CONCEPTS

• **Computing methodologies** → **Cognitive robotics**.

## KEYWORDS

Human-Robot Interaction, AI, Embeddings, Gaze, Multi-party

**ACM Reference Format:**
Léa Haefflinger, Frédéric Elisei, Brice Varini, and Gérard Bailly. 2024. Probing the Inductive Biases of a Gaze Model for Multi-party Interaction. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610978.3640702

## 1 INTRODUCTION

Even if gaze is one of the most studied non-verbal cues for Human-Robot Interaction (HRI) [1], gaze generation is still mostly restricted to highly controlled scenarios. The real-world implementation of natural gaze control for a robot raises a number of challenges, including its generalizability to new scenarios and new interlocutors. The control must be able to take into account and react effectively to a wide range of variabilities. We know, for example, that there are differences in the generation and perception of gaze depending on gender, age and culture [8, 12, 13, 18], but other variabilities likely have an effect. However, the majority of gaze controls proposed for multi-party interaction are heuristic models or are based on distributions obtained from human behavior studies [11, 15, 17]. With such predefined considerations, adaptability is not at the core of the built systems. In this paper, we propose to build a gaze management model for a robot in multi-party interaction, based on neural networks. By exploring the embeddings layers of this model, we aim to probe its inductive biases [6, 7], which could later be used as levers to drive a model that shows adaptation. Embeddings have already been successfully used in the speech domain [19], as well as in style transfer for virtual agent gesture generation [3]. We therefore believe that their use and exploration is an interesting approach in the context of robot gaze control.

This paper will first describe the construction of the gaze model for a game animator robot facing two players. Then, we'll focus on the analysis of its embeddings, such as that of a pair of players' bias, in order to understand the relationships that they may have captured. Finally, we'll check that the addition of player bias is used by the model and that this addition has a real impact on its performance.

## 2 MODEL DESCRIPTION

This section describes the construction of our robot's gaze model whose aim is to continuously generate gaze targets and meaningfull head movements [5].

### 2.1 Dataset for our Model

The dataset used is the RoboTrio2 corpus [4], a collaborative game with two human players, animated by an iCub robot [16] that is immersively teleoperated [2]. The robot transmits the voice of a human pilot located in an adjacent room and reproduces the pilot movements (eyes, head, lips). Wearing a virtual reality headset, the pilot can hear and see the two players through the robot's ears and its motorized eyes. The pilot's attention behaviour is recorded, along with the sound and two HD videos filming the players.


*HRI '24, March 11–14, 2024, Boulder, CO, USA*

ACM ISBN 979-8-4007-0323-2/24/03
https://doi.org/10.1145/3610978.3640702

In this game, the players have to find out which words are the most quoted from a theme word. While giving the themes and scores, the pilot has to encourage them to collaborate and provide answers. A tablet placed in front of the robot provides him with all the information about the game.

The dataset is composed of 11 sequences of around 20 minutes each (total of 4 hours), with 9 themes played in each. For each sequence, the robot's pilot, a man, is the same, but the player pairs are different and separated by gender, with male or female pairs.

## 2.2 Model Building

The gaze model proposed here (see Figure 1) is a cascaded model composed of causal Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) layers, where the first sub-model predicts gaze targets. This output is given as input to the second sub-model which predicts head movements. Gaze targets are limited to 5 classes: UserLeft (left player), UserRight (right player), Tablet, Elsewhere and OcularSaccade. For training data, the pilot's gaze has been automatically annotated using Gaussian models mixture (GMM). The predicted head movements correspond to the 3 recorded head angles: pitch (up/down), roll (tilt), yaw (left/right). To ensure good convergence of the head sub-model, it is trained with both predicted and real gaze targets. The models have a seq-to-seq architecture, with time sequences that correspond to a game theme (around 180 seconds per batch, 10 800 frames at 60Hz). The structure of the model is shown in Figure 1.

*2.2.1 Model Inputs and Biases.* The input features are multimodal, and are limited to signals and states that could be observed by a robot in a future real-time implementation. All inputs are passed through an embedding layer.

- **Endogenous to the robot (Pilot activity):**
  - *Speech:* speaking or not (embeddings dimension: 2)
  - *Speech Intent:* intent of the sentence, 24 different classes (asking for proposal, for validation, giving the theme, the score, ...) (embeddings dimension: 3). Classes were defined to best combine the various pilot sentences.
  - *Addressee:* 1 embedding for each player to cover the 0, 1 or 2 addressees cases (embeddings dimension: 2)
- **Exogenous to the robot (Users activity):**
  - *SpeechL, SpeechR:* whether left (resp. right) user is speaking or not (embeddings dimension: 2)
  - *GazeL, GazeR:* user is looking at the other user, at the robot, or at elsewhere (embeddings dimension: 2)

Features were annotated manually for verbal features, and automatically for users' gazes and the pilot's address [5]. In addition to these inputs, we decided to bias the model with the dyad number (the 11 pairs of players) facing the robot. We expect that the pilot's behavior can be influenced by certain social characteristics of the players. To encourage collaboration, he may need to adapt his behavior to balance participation if one player is more engaged than the other. This bias is encoded by an embedding of dimension 3.

*2.2.2 Model Training and Evaluation.* For the training of the model, two loss functions are used: categorical cross-entropy for gaze classification (5 targets) and Mean Square Error (MSE) for the prediction of the 3 head angles. The training process is separated into two stages, the first 50 epochs with a learning rate of 0.005, and the next 50 epochs with a learning rate of 0.001. In both stages, an Adam [10] optimizer is used. To best evaluate the model, we performed a 9-fold cross-validation. The mean F1-score for gaze classification is $0.57 \pm 0.01$ and the mean RMSE for head generation is $3.0° \pm 0.1$ with predicted gaze, and $2.1° \pm 0.1$ with real gaze.

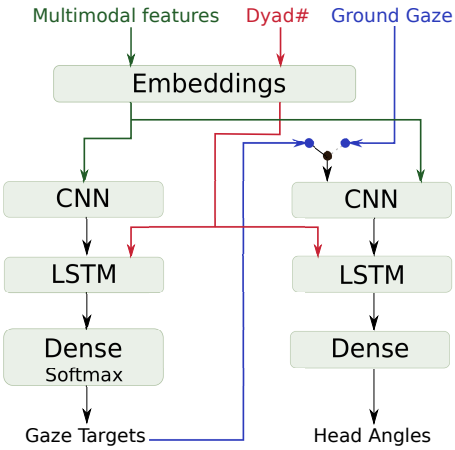


**Figure 1: Diagram of Model Structure.**

## 3 EMBEDDINGS ANALYSIS

To explore the inductive biases of our model, we performed an analysis of the embedding layers to identify what was encoded and what relationships could be inferred from the multimodal features. In this paper, we restrict analysis to the embedding matrix of the robot's *Speech Intent*, and that of the dyad number. The aim is to compare the representations of each class (each speech intention, each dyad) in the space of their embedding matrix and to identify if there are any correlations between these representations and social cues. To make the study of these correlations robust, they will be evaluated not on the embeddings of a single model, but on a combination of representations from our 9-fold models. The method used to obtain representations for *Speech Intent* and dyad embeddings is as follows: **1)** Reducing the dimensions of the embedding space, by applying a Principal Component Analysis (PCA) [20] to retain only the first two components (more than 90% of the explained variance). **2)** Each vector of the embedding matrix (1 vector per speech intention/dyad) is then projected into this reduced space. The result is a 2D representation of a set of 24 points for the *Speech Intent* embedding tensor, and respectively a set of 11 points for the dyad embedding tensor. **3)** Steps 1 and 2 are performed for the embeddings of the 9-folded models, giving 9 possible representations. In order to pool together these maps, the 9 sets of points obtained are realigned using the Kabsch-Umeyama algorithm to find the optimal translations and rotations [9]. As seed reference, we choose the set that minimises the final dispersion. The final representation of each speech intent/dyad number corresponds to the centroid of the covariance ellipse of the 9 realigned set of points.

The following sub-sections present the results obtained for the two analyzed features.

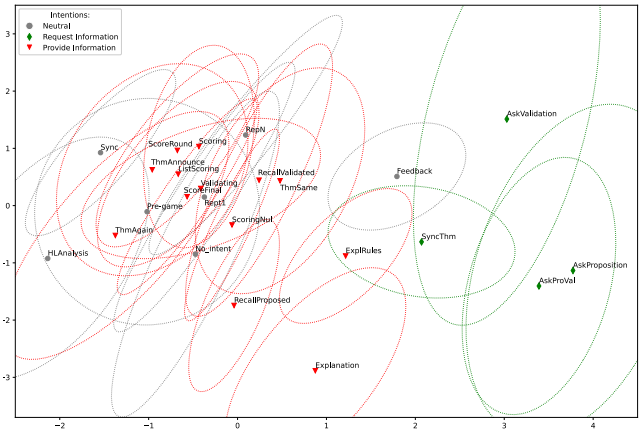


**Figure 2: Speech intents representation after realignement**

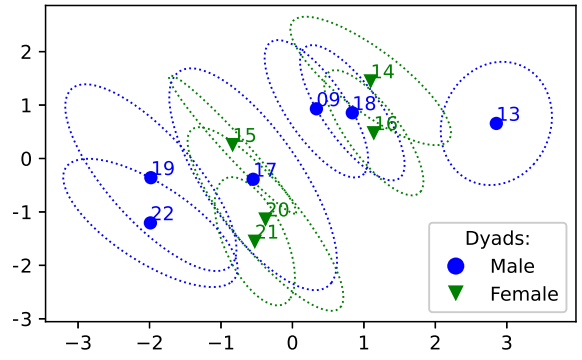


**Figure 3: Dyads representation after realignement**

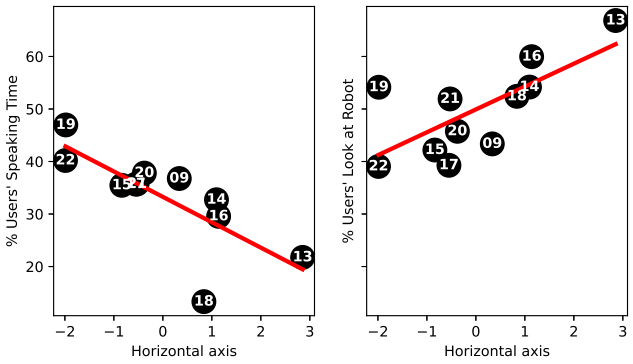


**Figure 4: Examples of linear regression between the horizontal axis of dyad embeddings representation and social cues**

## 3.1 Analysis of Embeddings of Speech Intents

Figure 2 shows the average representation obtained for all the speech intentions. The covariance ellipses are represented as well as their centroids. Confronted with this representation of the different speech intentions, we had the intuition that two groups emerge: either by providing or by requesting information. To confirm this first impression, we divided the intentions into three groups and assigned a colour to each. Intentions classified as providing information are represented in red with a triangle marker, those classified as requesting information in green with a diamond marker, and neutral intentions in grey with a round marker. We can therefore see that the horizontal axis acts as a scale for the transfer of information: with on the left the speech intents that provide information (giving the theme, the score, etc.) and on the right those that request information (asking for proposals, for validation of proposals, or if people are ready for the next theme with SyncThm).

The clustering of speech intents by the model seems consistent with the communicative intentions they convey. This could be partly explained by the need to use (or not) the tablet during certain utterances, and therefore to look at it to read the information it contains. This is a particularly interesting observation, which could lead to the use of this model in new interaction scenarios, where the speech intents could be different, but placed in relation to those already employed.

## 3.2 Analysis of Embeddings of Dyads of Players

Figure 3 shows the average representation obtained for all dyads with their covariance ellipse and centroid. The first hypothesis we wanted to verify was that of a grouping based on the gender, i.e. is there a relationship between the gender of the players in the dyad and their representation in the embedding space? For this, the points of dyads composed of female players are represented in green with a triangle marker, and those composed of male players in blue with a round marker. No grouping can be made on the basis of gender. The embedding layers do not appear to have encoded player gender information. If 3 groupings seem to emerge (22, 19, 20, 21, 15, 17), (09, 18, 14, 16), and the isolated 13, it is difficult to explain them simply by looking at this figure. To find any hidden relationships behind these clusters, we calculated correlations between the representation axes and quantifiable social indices. Table 1 summarizes the correlations obtained after a Spearman test [14].

The horizontal axis seems to be strongly correlated with two social cues: the percentage of time during the interaction when at least one of the two players was speaking, and the percentage of time when at least one of the players was looking at the robot. These two social cues can be linked to a measure of the players engagement

**Table 1: Correlation coefficients between social cues and the two axis of dyad embeddings representations**

| Social Cues | Horizontal Axis | | Vertical Axis | |
|---|---|---|---|---|
| | Corr | p-value | Corr | p-value |
| Final Game Score | -0.42 | 0.20 | -0.3 | 0.37 |
| Users' Mean Age | 0.1 | 0.75 | -0.18 | 0.6 |
| Diff Users' Age | -0.02 | 0.96 | 0.15 | 0.65 |
| %Users' Speaking Time | **-0.76** | **0.006** | -0.57 | 0.06 |
| Diff Users Speaking Time | 0.23 | 0.50 | 0.27 | 0.42 |
| %Users Mutual Gaze | -0.42 | 0.20 | -0.47 | 0.14 |
| %Users Look at Robot | **0.76** | **0.006** | 0.48 | 0.13 |

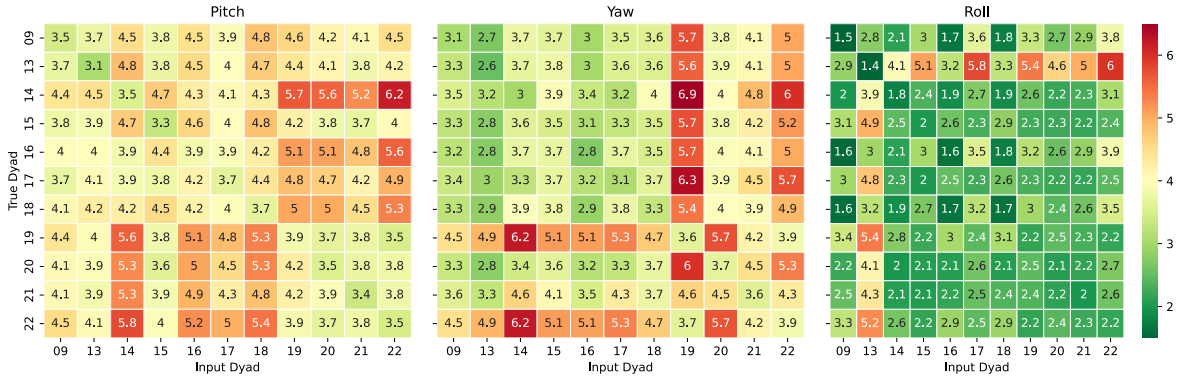


**Figure 5: Heatmap of RMSE performances for each head angle depending on the dyad number given as input**

in the interaction. However, the correlations are opposite, with an inversely proportional correlation for speech percentage, which is not the case for gaze percentage. Figure 4 shows the linear regressions obtained for these two correlations (on left with percentage of speech, on right percentage of gaze), where we can clearly see this opposition. One hypothesis to explain this opposition could be that very active players will think a lot and talk more to each other to make proposals, and therefore look less at the robot than more passive players. This result is particularly interesting, and seems to provide an inductive bias for actively adapting the model according to the players' involvement in the interaction, which is a necessity for a robot animator.

No strong correlation was found on the vertical axis. We haven't tested every possible social cue. Furthermore, this analysis is based on only 11 points, and is limited to the search for linear correlations, which is not necessarily the case. Note that "causal" regression is a method that attempts to explain what our model might have captured, but cannot prove causality.

## 4 IMPACT OF EMBEDDINGS OF DYADS ON GAZE AND HEAD PREDICTIONS

In the previous section, we highlighted certain biases in our model, especially linked to player behavior. Due to these biases, the model clustered dyads that were more or less similar, but do these differences really impact predictions? We then tested the predictions of our model by swapping the dyad number given as input. In total, an interaction sequence with dyad number N will be predicted 11 times, once by giving the dyad number N as input, and 10 times by giving the numbers of the 10 other dyads. Each prediction is then evaluated, by calculating an F1-score for the gaze targets, and the RMSE value for each of the 3 head angles.

Figure 5 shows heat maps of head angles generation performance for each dyad. The diagonal of each heat map corresponds to the performance when no swap has been applied. We can confirm that the best performances for each dyad are obtained on this diagonal. We can clearly see that the model's performance can vary considerably depending on which of the dyad numbers is swapped. In more detail, we can see that for the Pitch angle, dyads numbers 14, 16, 17, 18 have their results strongly impacted when we replace their number with dyads 19, 20, 21, 22, with differences close to 2 degrees, and reciprocally. For the Yaw angle, it's dyads 19 and 22 that seem to distinguish from the 9 others: on Figure 3, we could see that these two dyads were positioned close together and a little more in the negative. Finally, on the Roll angle heat map, the lines of dyad 13 stand out clearly, with performance differences of almost 5 degrees, this dyad is actually a bit isolated on the Figure 3.

Regarding the impact on gaze target prediction performance, no dyad appears to differ clearly from the others, but performance can drop to F1-scores below 0.45 for some swaps, well below the average at 0.57.

The dyad number bias has a real impact on the performance of our model. The importance of this impact can be associated with the representation of the dyads in the embedding spaces. Swapping two dyad numbers that are distant from each other in the map induces a significant drop in performance, whereas swapping dyads that are close to each other results in a smaller drop in performance.

## 5 CONCLUSIONS

To design a gaze model for our robot interacting with two people that could be easily generalised to other scenarios and other interlocutors, we proposed using neural networks biased by inductive embeddings. By analysing the spaces of these embeddings, we were able to identify some possible inductive biases that could be used for generalisation. We observed that the model made a distinction between speech intentions that corresponded to requests for information and those that provided information. Similarly, it groups together different pairs of interaction partners, which could be explained by social cues linked to engagement such as speaking time and gaze. We then show that these distinctions have a real impact on the model's predictions. These results are promising and seem to offer a solution for better flexibility and adaptation of the model to be reused in front of different pairs of people and for other scenarios of interaction such as a hotel receptionist robot or a bartender. Such inductive bias can only be found as a by-product of end-to-end modelling: this analysis-by-generation process hopefully contributes in bridging science and deep learning technology.

## ACKNOWLEDGMENTS

This work is supported by the ANR 19-P3IA-0003 MIAI. The first author is financed by a CIFRE PhD granted by ANRT (2021/0836).

# REFERENCES


[1] Henny Admoni and Brian Scassellati. 2017. Social Eye Gaze in Human-Robot Interaction: A Review. *J. Hum.-Robot Interact.* 6, 1 (may 2017), 25–63.

[2] Remi Cambuzat, Frédéric Elisei, Gérard Bailly, Olivier Simonin, and Anne Spalanzani. 2018. Immersive Teleoperation of the Eye Gaze of Social Robots Assessing Gaze-Contingent Control of Vergence, Yaw and Pitch of Robotic Eyes. In *ISR 2018 - 50th International Symposium on Robotics*. VDE, Munich, Germany, 232–239.

[3] Mireille Fares, Catherine Pelachaud, and Nicolas Obin. 2023. Zero-shot style transfer for gesture animation driven by text and speech using adversarial disentanglement of multimodal style encoding. *Frontiers in Artificial Intelligence* 6 (2023), 1142997.

[4] Laurent Prévot Frédéric Elisei, Gérard Bailly. 2023. RoboTrio2. https://hdl.handle.net/11403/robotrio/v2 ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

[5] Léa Haefflinger, Frédéric Elisei, Silvain Gerber, Béatrice Bouchot, Jean-Philippe Vigne, and Gérard Bailly. 2023. On the Benefit of Independent Control of Head and Eye Movements of a Social Robot for Multiparty Human-Robot Interaction. In *Human-Computer Interaction*, Masaaki Kurosu and Ayako Hashizume (Eds.). Springer Nature Switzerland, Cham, 450–466.

[6] Thilo Hagendorff and Sarah Fabi. 2023. Why we need biased ai: How including cognitive biases can enhance ai systems. *Journal of Experimental & Theoretical Artificial Intelligence* (2023), 1–14.

[7] Eyke Hüllermeier, Thomas Fober, and Marco Mernberger. 2013. *Inductive Bias*. Springer New York, New York, NY, 1018–1018.

[8] Carlos Toshinori Ishi and Taiken Shintani. 2021. Analysis of Eye Gaze Reasons and Gaze Aversions During Three-Party Conversations. In *Proc. Interspeech 2021*. 1972–1976.

[9] Wolfgang Kabsch. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* 32, 5 (1976), 922–923.

[10] Diederik Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (2014).

[11] Chinmaya Mishra and Gabriel Skantze. 2022. Knowing Where to Look: A Planning-based Architecture to Automate the Gaze Behavior of Social Robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1201–1208.

[12] Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. 2006. A storytelling robot: Modeling and evaluation of human-like gaze behavior. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 518–523.

[13] Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. 2012. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems* 1 (01 2012), 12.

[14] Leann Myers and Maria J Sirois. 2004. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences* 12 (2004).

[15] Yukiko I. Nakano, Takashi Yoshino, Misato Yatsushiro, and Yutaka Takase. 2015. Generating Robot Gaze on the Basis of Participation Roles and Dominance Estimation in Multiparty Interaction. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 22 (dec 2015), 23 pages.

[16] Alberto Parmiggiani, Marco Randazzo, Marco Maggiali, Frederic Elisei, Gerard Bailly, and Giorgio Metta. 2014. An articulated talking face for the iCub. In *2014 IEEE-RAS International Conference on Humanoid Robots*. 1–6. https://doi.org/10.1109/HUMANOIDS.2014.7041309

[17] Taiken Shintani, Carlos T. Ishi, and Hiroshi Ishiguro. 2021. Analysis of Role-Based Gaze Behaviors and Gaze Aversions, and Implementation of Robot's Gaze Control for Multi-Party Dialogue. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (Virtual Event, Japan) *(HAI '21)*. Association for Computing Machinery, New York, NY, USA, 332–336.

[18] Gabriel Skantze. 2017. Predicting and Regulating Participation Equality in Human-Robot Conversations: Effects of Age and Gender. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 196–204.

[19] Shirui Wang, Wenan Zhou, and Chao Jiang. 2020. A survey of word embeddings based on deep learning. *Computing* 102 (2020), 717–740.

[20] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.