People identity learning in HRI through an incremental multimodal approach

Soumaya Sabry*, Mohamad Ghassany*, and Salvatore M. Anzalone†
*Léonard de Vinci Pôle Universitaire, Research Center, Paris La Défense, France
†Laboratoire CHArt, Université Paris 8, Saint-Denis, France

***Abstract*—In order to achieve a real partnership with humans, robots should be able to personalize and adapt their behaviors to the specifity of each human partner. A fundamental step to achieve this goal is to develop a reliable characterisation of their identity. In this paper, we present a robotic system capable of learning the identities of humans through an incremental multimodal approach. Features from faces and voices are extracted as people signature and exploited to achieve a self-supervised learning. A feasibility study helped us to underline challenges, opportunities and possible obstacles in the development of such skills.**

***Index Terms*—Social robotics, people identity, multimodal integration, incremental learning, face recognition, speaker identification.**

I. Introduction

In recent years, research efforts focused on giving robots the appropriate cognitive capabilities to autonomously act in a safe and intelligent way in real world scenarios [1].

In fact, the large majority of robots are used in controlled spaces with a known structure that sometimes is also modeled around the requirements of the robot itself [2]. In such scenarios, robots with a complete knowledge of their environment can be programmed to maximize their performances in terms of speed and precision [3]. However, while such robots can be extremely useful in industries and in factories for various activities, ranging from painting to picking and placing objects, many "real world" scenarios are left out. In fact, in semi-structured or non-structured, dynamic environments, robots can have only a partial control and a partial knowledge of the world. To deal with them, they need cognitive abilities to perceive, reason and act to unexpected conditions, in continuously changing environments [4]. Moreover, in a wide range of activities, robots endowed with social abilities could behave as real, effective partners in both cooperative [5] and competitive scenarios [6]. Social robotics focuses on the grand challenge of endowing robots with sociocognitive skills that take in an explicit account the human presence in their perception-decision-action loop[7]. This entails the integration of various social skills, ranging from a fine perception and characterization of the human presence (people detection, emotion recognition, activity identification), to the coherent production of "believable" behaviors [8].

Corresponding author: S. M. Anzalone (email: sanzalone@univ-paris8.fr) and M. Ghassany (email: mohamad.ghassany@devinci.fr).

However, in order to achieve a real partnership between humans and robots, sociocognitive skills are not enough: humans are individuals and want to be treated as such [9]. Robots should be able to personalize and adapt their behaviors to the specificity of each human partner. Such personal robots, would tailor their interactive skills (behaviors, lexical register, etc) to the needs of their partners, anticipating them, taking in account their personal preferences, providing customized help and suggestions.

The first, fundamental step to produce a robot's internal model of human partners is to develop a reliable characterisation of their identity based on their physical characterisation [10]. In this paper we propose an incremental identity learning system based on a self-supervised model built through the integration of synchronized multimodal features. We focus, in particular, on faces and voices of robot's partners in in dyadic human-robot interaction (HRI) scenario. In the following we will introduce state of art, our method, some implementation details and the results of a feasibility study that helped us to underline challenges, opportunities and possible obstacles in the development of such skills.

II. State of art

The system presented here allows the multimodal identification of human partners. Several attempts to the person identification problem have been achieved. Ishiguro et al. [11] presented a multi-robot system "MINERVA" working in a museal environment in which people was identified and tracked using RFID. This solution is effective but it needs that all the users bring with them this kind of identifiers.

According to a cognitive approach, other solutions have been developed to make robots capable to identify humans through their biometric features [12]. Face identification systems and voice identification systems

have been used in these robots to achieve this purpose. These robots were used to target a special kind of service designed for personal use at home, they are the so-called robot companions or social robots.

Murray and Canamero developed Erwin [13], a robot capable to learn to recognize its caregivers through the use of voices features and faces features. ERWIN's system learns faces and voices by interacting with partners during a first learning stage. After that, the robot is able to recognize them and showing them an emotional expression generated by its system. ERWIN's face recognition system uses a face extraction method that generates a vector of five features: distance between eyes, mouth width, distance between eye level and mouth, distance between eye level, and centre of nose. As drawback, the accuracy of the face recognition decreases with any head rotation of the person (in case of a rotation of 30 degree, the system gives 0% correct value).

All the presented works underline the importance of a correct binding between the representations coming from different modalities. This is the problem of "anchoring", as defined by Coradeschi and Saffiotti [14], [15]. Recent studies from cognitive sciences [16] show that facilitation and interference effects are associated to combinations of visual and vocal features: a correct anchoring between voices and faces should facilitate the learning of people identities. Then, binding modalities becomes one of the main problems in the multimodal approach for identifying the robot's partner.

Here we propose a multimodal approach capable of learning the identities by combining visual and vocal modalities. In a recent work, a multimodal approach has been used in a work focused on helping elderly people [17] through an assistive robotic platform that exploits multiple sensors for human action recognition.

A more recent article [18] presents a multimodal Bayesian network for person recognition in a HRI context. Authors propose a model to identify a person based on her/his face, gender, age and height estimates. Human user will guide, through feedbacks on a tablet, a supervised identity learning system. As contrary to this approach, we propose the use of multimodality, face features and voice features, to enforce a self-supervised identity learning.

A similar system to our approach [19] is used in security purposes. It uses face and speech recognition to identify the client and recognise him by referring to their database. As for face extraction, it uses Eigenspace modeling to define the probability of a match between a person and eigen-coefficients given the person's model. As for speech recognition, it uses MFCs [20] and modeling by HMM parameters [21]. The system builds a classifier for each modality and a confidence score for each classifier. To combine both modalities they use a Bayes Net [22]. Their results show high accuracy for identifying the right customer using an ATM. This approach is supervised and needs a labeled database to which we can refer, in contrary to the approach we propose in this paper for which no prior knowledge is required.

In this paper, we present a multimodal approach for learning the identities. We use a *self-supervised learning* technique, during which a clustering technique is applied on the face modality, then combined with a classification technique on the voice modality. The classification phase uses the cluster labels to build its prediction model. In the rest of the paper, we present a state of the art of the face and voice recognition techniques, then we explain our approach and discuss the results of our simulation.

III. System Overview

The aim of our project is the development of a robotic system capable of learning the identity of humans through a multimodal approach. Voices and faces data from the environment are perceived through the microphone and the camera of the robot. This raw information is analyzed to detect humans features that, if any, are extracted and collected. The features collected represent a biological signature of each person, so they are opportunely treated in a self-supervised way to obtain identification claims.

Our system pipeline is divided into four phases, it is described in Fig. 1. The four phases are: data collect, data processing, data wrangling, modeling and prediction. We describe the four phases in this section and we present a more detailed overview of some phases in the next section.

In the first phase data is collected separately. Starting with an Audio/Video stream (mp4, live stream, recording, etc), each modality is first separated to be treated simultaneously in the second phase. During the second phase, the collected data is processed and treated according to its nature, audio or video. Concerning the visual modality, a face detection algorithm is first applied, if any faces are found, and then the face features are extracted. In the other hand, a voice activity system distinguishes voices from noise, then, eventual voice features are extracted.

Since the audio sampling rate is bigger than the image sampling rate, each processing method outputs a different number of rows. Consequently, a downsampling and a synchronization is needed. In this third phase, that we call *data wrangling*, we bind each face frame to multiple voice frames according to their timestamp. Moreover, to minimise errors coming from noisy inputs, we treat data as blocks containing a fixed number of rows. Blocks are particularly useful to detect new acquaintances, as the system will have enough information to create a new cluster associated to the novelty. Therefore, each

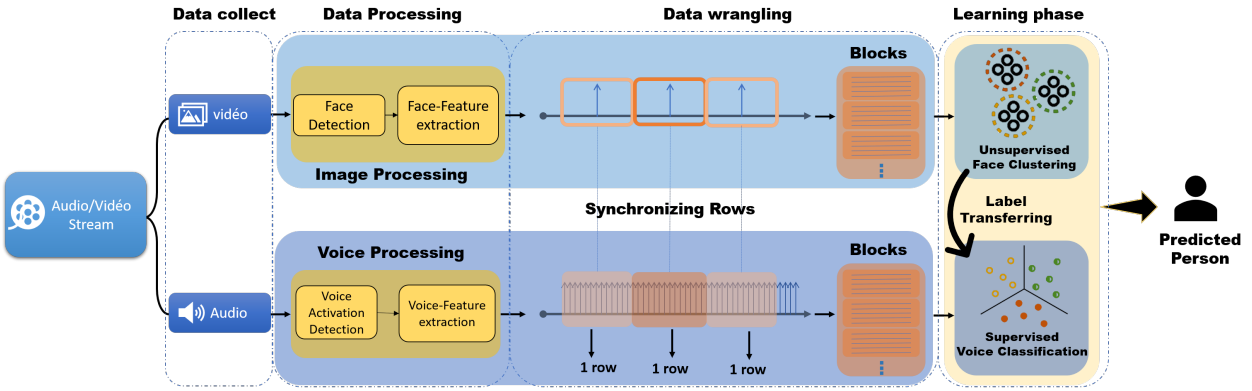

Fig. 1: The proposed system pipeline.

block becomes a basic computational units that will be exploited by the proposed algorithm.

The goal of the system is to learn and to predict the identity of the robot's partner, without any prior knowledge or pre-trained model. A possible way to achieve this result is to apply an unsupervised clustering technique on both of the data sets (face and voice). But while this could be feasible with face features, audio features present many difficulties for clustering because of their noisiness, thus the difficulty of detecting well separated clusters. Beside that, merging the two datasets before the clustering does not allow a recognition of the identity using only one modality. At the same time, if we simply merge the two datasets, one of the modalities, the faces, dominates the other one. This is logical in a way, as the face features are more distinguishable than voice features, and the information collected from the voice source becomes useless in this case [23].

We propose, then, to use a self-supervised approach: first, face features are clustered, then the obtained cluster labels are transferred to the voices features data set. A supervised classification technique is then applied on these features using the transferred labels. During this leaning phase a model is obtained and its performance are evaluated continuously. The performance of the model is evaluated, here, by calculating the accuracy by blocks. Note that, since we do not have any prior knowledge about the number of identities that should be recognized, we determine the optimal number of clusters with a cluster validation technique. At the end of the learning phase, a model is obtained with labels corresponding to the identity of the persons.

In the last phase, when a new person is presented, the robot will be capable of using its own produced and learned model to predict the person's identity. If this person is new, it will increment the set of known acquaintances, creating a new cluster label and retrain the model.

IV. Implementation Details

This system can be realized with various techniques for each phase. After testing multiple techniques, we decided to apply the HOG technique for face detection since we noticed that it is more accurate than Haar-cascade and faster than CNN detection. Moreover the face features are retrieved through Face encoder (deep-learning trained model) based on Face Net [24]. From a short time spectral analysis of the voice samples, the Mel Frequency Cepstral Coefficients (MFCC) are extracted and used as voice features. In the rest of this section we describe the techniques we used in our approach.

A. Face data processing

The HOG technique [25] is used for detecting faces in an image. The input of this method is an image in grayscale. The main idea is to describe local object appearance and shape by the distribution of local intensity gradients or edge directions, even without accurate knowledge of the corresponding gradient or edge positions.

As well described by [25], the HOG method divides the image window into small spatial regions ("cells"). For each cell, it collects a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. The combined histogram entries form the representation of a generic face. In details, the images are split into squares of 16×16 pixels. The method counts how many times each direction has been discovered before, and only one arrow is drawn in the square according to the most frequently found direction. This action is performed for each square of pixels in the image.

Once the detector has been trained, it is ready for use. For detecting a face, the detector will go through the image, search for a face and check if the pattern matches somewhere. Each time the pattern is identified somewhere in the picture, it means a face has been detected.

After the detection of a face, instead of engineering features [26], we use a pre-trained Deep learning Convolutional Neural Network to generate a 128 features [24] space representing the characteristics of faces of individuals. The network has been trained by OpenFace[1] and implemented through the Python library "Face Recognition" [2].

Notably, the pre-trained neural network autonomously adjusted its parameters during its training (Fig.2) to make sure that the features generated for similar faces would be closer between them, while distant from other faces. This results on a space that encodes identities, separating in a reliable way the facial features from each person.

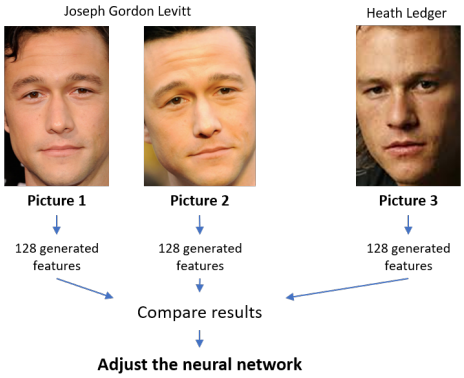

Fig. 2: The pre-trained CNN adjusted its parameters to produce face features from Pictures 1 and 2 closer to each other while distant from those of Picture 3.

B. Voice data processing

Voices are detected through a Voice Activity Detection algorithm (VAD) [27], based on the thresholding of the signal energy, interpreting high-energy regions as speech. If the energy of the signal is above an empirically defined threshold, the VAD indicates speech activity.

As for voice features, Mel-Frequency Cepstral Coefficients (MFCC) are common and widespread features used in speech analysis and speech recognition. Such features are calculated in the frequency domain according to the Mel scale which is based on the human ear scale. More in detail, the audio signal is split into time frames containing an arbitrary number of samples. Overlapping of the frames is considered sometimes into smooth transition from frame to frame. After that, each time frame is windowed with Hamming window [28] to eliminate discontinuities at the edges. Then, Fast Fourier Transformation (FFT) is calculated for each frame to extract frequency components of a signal in the time-domain. A logarithmic Mel-Scaled filter bank is then applied to the Fourier transformed frame. This scale is approximately linear up to 1 kHz and logarithmic at greater frequencies. The relation between frequency of speech and Mel scale [29] can be expressed as:

$$mel(f) = 2595 \log_{10}(1 + f/700)$$

The resulting bank of filters have greater bandwidths for higher frequency filters than for lower frequency filters. Finally, Discrete Cosine Transformations (DCT) of the logarithmic output vectors from the filters banks are calculated. The overall process of MFCC extraction is shown on Fig. 3.

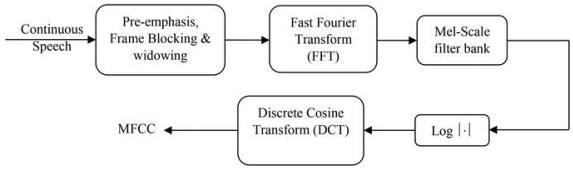

Fig. 3: The MFCC extraction system as presented in [30]

C. Multimodal learning

To treat the two synchronized modalities, our algorithm enforces a master-slave relationship between them. More reliable than the sound, the video modality is chosen as master modality: faces predictions, consequently, will train the voices model. While this assure the possibility of using a single modality to recognize people, the performance of the voice model will depend strongly on the performances of the face model. As drawback, the recognized face will always has priority on voice information; at the same time, if a face is wrongly predicted, the voice features would be wrongly trained.

Each buffered block dispatched to the learning subsystem contains new, unseen multimodal features. Running

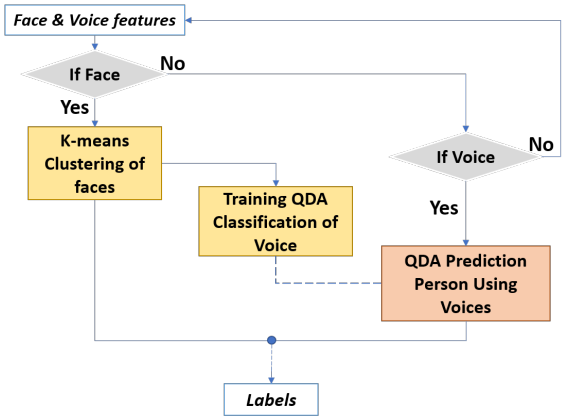

Fig. 4: Multimodal learning decision flow

[1] https://cmusatyalab.github.io/openface/
[2] https://github.com/ageitgey/face_recognition

through it, as in Fig. 4, the system verifies if any face was found. In the positive case, it starts the process of clustering. New acquaintances are handled by the system using the silhouette score [31] to get the optimal number of clusters. In a second step, the optimal number is passed to the k-Means clustering algorithm to obtain the predicted label of the person seen. Once retrieved, this label is transferred to the audio modality to train a voice model from the synchronized audio data, using a Quadrative Discriminant Analysis (QDA) classifier. In the case in which no face was detected, the system checks the VDA: if a voice is found, the system will rely on the QDA trained voice model to recognize it.

In order to improve the performances of the system, wrong identifications are removed through an absolute frequency analysis of the labels found in each buffered block: the most recurrent one was interpreted as the predicted person for that block.

V. Experiments

Working with a real robotic platform translates multiple issues coming from the inner complexity of operating in the real world: real-time capabilities, multiparty interaction, cocktail party effect, echos, environmental noise, the noise of the robot itself due to its motors, etc. To relax such constraints, we use a simulated approach, by employing a database of multiple video streams.

The algorithm simulates the robot's perceptions using raw data from a dataset of audio/video files (mp4 file format). Each video is supplied to the proposed system and exploited frame-by-frame, like described earlier. In the following, we describe the employed dataset and we present a detailed quantitative analysis of the evolution of the accuracy: using single modalities or by employing our multimodal algorithm.

A. Dataset

In order to test and verify the effectiveness of the presented approach, we created a database composed by videos of bloggers. This choice has been made because the interaction the bloggers have with their followers could be perceptually similar to a direct, face-to-face interaction with a robot. The blogger, in fact, talks directly to them through the camera, exploiting a wide range of non-verbal cues as gestures, gazing, facial expressions. From the perceptive point of view of the robot, this could replicate a dyadic HRI scenario in which a human interacts directly to a robot, explaining or giving instructions to a particular task. At the same time, such videos are characterized by their high quality and by a clear and understandable voice. An important assumption made during the development of this system is that speakers will not overlap their voices and voices heard come only from persons shown on screen.

The dataset is composed by two male bloggers[3] and one female blogger[4] to highlight sex differences. Notably, one of the male bloggers speak Arabic, while the others speak English. The database employed was composed by 4 videos, divided in a total of 8 chunks, concatenated randomly into a single MP4 file for a total of 4.5 minutes of data (Fig. 5). Its data sampling rate was of 25 fps for the video and 44100Hz for the audio.

Fig. 5: Video samples from the exploited dataset.

In order to reduce bias, the dataset was kept artificially balanced by carefully selecting the amount of time that each speaker spent in front of the camera.

B. Results

Several experiments have been performed in order to verify the effectiveness of the system.

1) Face clustering performances: Fig. 6 represents the temporal evolution of the incoming faces features projected through PCA into a bi-dimensional space. Three persons appear one after another. Each color represents a different person. Incoming face features form incrementally several clusters, one for each face. The system starts in (1) with a limited amount of data coming from a single person; thus, the clustering algorithm associates this data to a single cluster, giving it an unique identifier. As long as new data is processed by the system, the cluster associated to a first identifier grows up (2-3). The appearance of a second person will produce and nourish a new cluster. This will be correctly identified as a new acquaintance through the analysis of the silhouette score (4-6). The process occurs again in a similar way for a third person (7-9). The arriving of an already known acquaintance will be correctly inferred by the system: features of already known people, in fact, will nourish existent clusters, without causing any considerable change on the silhouette score.

Fig. 7 shows the evolution in time of the accuracy of the face clustering system, per data block. While, in general, the system achieves high accuracy rates, in some specific blocks such rates fall down, dropping sometimes at zero, as the system was not able at all to perceive any face. Tracing back the causes, we identified 3 main motivations:

[3]NasDaily,https://www.youtube.com/user/nyassin14 and Hazem el Seddiq, https://www.youtube.com/user/Hazmaniac1

[4]Amanda RachLee, https://www.youtube.com/user/amandarachlee

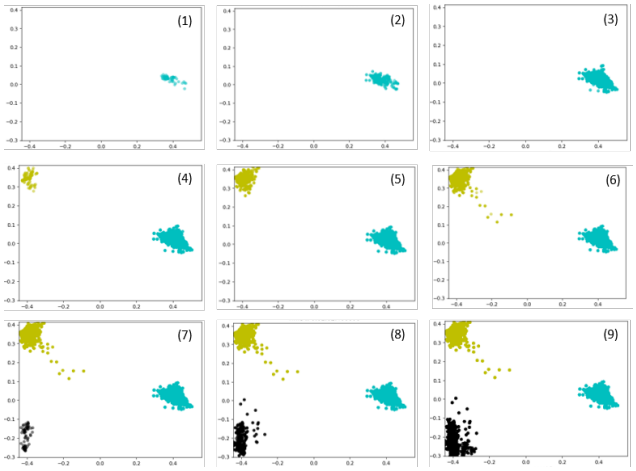

Fig. 6: Temporal evolution of the incoming faces features projected into a bi-dimensional PCA space. Each color represents a different person.

(a) Occlusions: due to the extensive use non-verbal language and gestures, the blogger sometimes puts the hands in front of her/his face;
(b) Out-of-bounds head rotations: the head of the blogger is rotated out of the detection bounds of the face detection system;
(c) Turn taking: during a switch between people, a different person arrives in front of the camera.

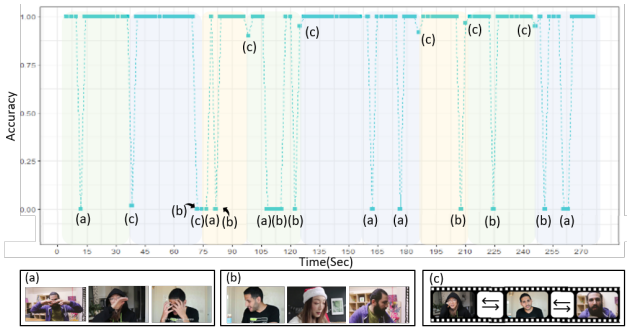

Fig. 7: Face model accuracy performances in time annotated with eventual failure causes: (a) occlusions; (b) out-of-bounds head rotations; (c) turn taking.

2) Voice classification performances: Fig. 8 shows the evolution in time of the accuracy of the voice classification system, per data block. Predicted labels come from the trained classification voice model. Also in this case, despite its general high performances, the audio recognition system presents in some data blocks consistent drops on performances. Causes can be tracked to the following:

(a) New acquaintances: a voice of a new person is perceived for the first time without seeing the correspondent face; consequently, the face model has still not been trained and the voice would belongs to an yet unlabeled person;
(b) Noisy perception: the voice is not clear or environmental noise is present;
(c) Turn taking: during a switch between people, a different person starts to talk.

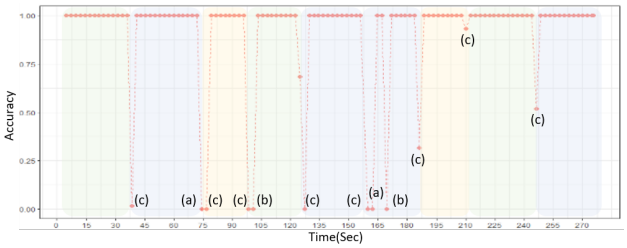

Fig. 8: Voice model accuracy performances in time annotated with eventual failure causes: (a) new acquaintances; (b) noisy perception; (c) turn taking.

3) Multimodal learning performances: Fig. 9 shows the evolution per data block in time of the accuracy of the proposed multimodal identity learning system (red, audio features based accuracy; blue, video features based accuracy; green, multimodal system accuracy).

The figure clearly shows how the accuracy of the multimodal system outperforms the performances of the single modalities, recovering their failures. Despite of this, the system still presents some limitations coming mainly from situations in which the face is not detected and the quality of audio is bad, or, as in the cases of single modalities, during turn taking, when a person follows a different one.

VI. Discussion

This paper has not the ambition of confirming the effectiveness or the efficacy of the proposed system: the dimension of the proposed dataset, in fact, precludes any generalization of the obtained results. The performed experiments using simulated data, however, gave us the opportunity of testing the concept of an incremental identity learning system that achieves a self-supervision through the integration of synchronized multimodal features. The tests accomplished show the feasibility of the proposed approach, highlighting its limits and its potentialities, as well as its perspectives in terms of technological development and of possible issues. The presented results are preliminaries and limited by several factors:

- Offline execution: due to its inner nature of prototype, the presented system runs offline, without the application of any temporal restrictions. Real-time constraints will be enforced for the implementation of the system in a robotic platform;

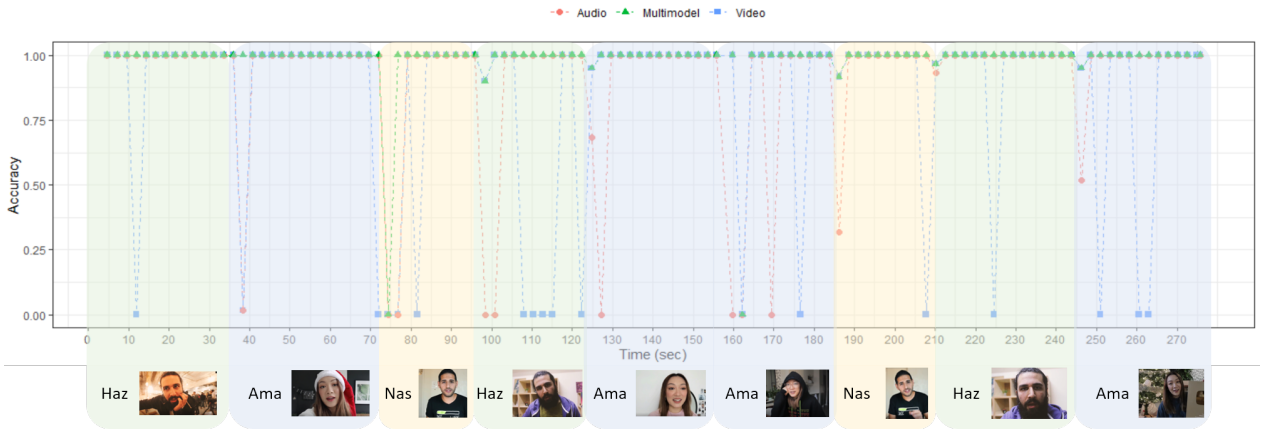

Fig. 9: Temporal evolution per data block of the accuracy of the proposed identity learning system exploiting single modalities (faces and voices) and their integration.

- Focus on dyads: the current demonstrator focuses on direct, face-to-face communication of a human with a robot. A real system should be able to deal with multiple people present in the same scenario;
- Single speakers: the developed prototype assumes that the perceived voice can come only from the single person in front of the camera. An embodied system in a real world scenario should be able to deal with multiple voices;
- Ignored spatial information: the presented system assumes that people would always appear inside the robot's field of view. A complete system should deal with people present in the environment but outside the robot's field of view;
- High-quality devices: the prototype developed has been tested using high quality videos while robots' on-board sensors are usually of a lower quality. A functional embodied system should be able to deal with such kind of unreliable data;
- Quiet environment: the test dataset employed is composed by videos in which voices are clear and loud, while real environment are usually noisy. Robots should be able to filter out any source of noise as sound from external sources, overlapped voices, being able to focus its own auditory attention (cocktail party effect).

Future efforts will focus in tackling these issues. Possible solutions will comprehend the integration in the current system of a wide range of technologies able to localise and track data sources while filtering noise.

Despite such limits, this paper shows the potential of the proposed approach applied to the identity learning problem. Future works will focus on developing and evaluating a possible implementation on a real robot, in real world scenarios.

REFERENCES

[1] F. Ingrand and M. Ghallab, "Deliberation for autonomous robots: A survey," Artificial Intelligence, vol. 247, pp. 10–44, 2017.

[2] V. Villani, F. Pini, F. Leali, and C. Secchi, "Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications," Mechatronics, vol. 55.

[3] B. Siciliano and O. Khatib, Springer handbook of robotics. Springer, 2016.

[4] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori et al., "Integration of action and language knowledge: A roadmap for developmental robotics," IEEE Transactions on Autonomous Mental Development, vol. 2, no. 3, pp. 167–195, 2010.

[5] R. Bemelmans, G. J. Gelderblom, P. Jonker, and L. De Witte, "Socially assistive robots in elderly care: A systematic review into effects and effectiveness," Journal of the American Medical Directors Association, vol. 13, no. 2, pp. 114–120, 2012.

[6] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, and E. Osawa, "Robocup: The robot world cup initiative," in Proceedings of the first international conference on Autonomous agents.

[7] C. Breazeal, "Toward sociable robots," Robotics and autonomous systems, vol. 42, no. 3-4, pp. 167–175, 2003.

[8] C. Breazeal, Designing Sociable Robots. Cambridge, MA, USA: MIT Press, 2002.

[9] S. M. Anzalone, Y. Yoshikawa, H. Ishiguro, E. Menegatti, E. Pagello, and R. Sorbello, "Towards partners profiling in human robot interaction contexts," in International Conference on Simulation, Modeling, and Programming for Autonomous Robots. Springer, 2012, pp. 4–15.

[10] S. M. Anzalone, E. Menegatti, E. Pagello, Y. Yoshikawa, H. Ishiguro, and A. Chella, "Audio-video people recognition system for an intelligent environment," in 2011 4th International Conference on Human System Interactions, HSI 2011.

[11] S. Thrun, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Schulte et al., "Minerva: A second-generation museum tour-guide robot," in Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C), vol. 3. IEEE, 1999.

[12] A. Jain, R. Bolle, and S. Pankanti, "Introduction to biometrics," in Biometrics. Springer, 1996, pp. 1–41.

[13] J. C. Murray and L. Cañamero, "Developing preferential attention to a speaker: A robot learning to recognise its carer," in 2009 IEEE Symposium on Artificial Life. IEEE, 2009, pp. 77–84.

[14] S. Coradeschi and A. Saffiotti, "An introduction to the anchoring problem," Robotics and autonomous systems, vol. 43, no. 2-3, pp. 85–96, 2003.

[15] S. Coradeschi, H. Ishiguro, M. Asada, S. C. Shapiro, M. Thielscher, C. Breazeal, M. J. Mataric, and H. Ishida, "Human-inspired robots," IEEE Intelligent Systems, vol. 21, no. 4, pp. 74–85, 2006.

[16] S. Campanella and P. Belin, "Integrating face and voice in person perception," Trends in cognitive sciences, vol. 11, no. 12, pp. 535–543, 2007.

[17] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos, "Multimodal human action recognition in assistive human-robot interaction," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 2702–2706.

[18] B. Irfan, N. Lyubova, M. G. Ortiz, and T. Belpaeme, "Multimodal open-set person identification in hri," in 2018 HRI Workshop, 2018.

[19] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland, "Multimodal person recognition using unconstrained audio and video," in Proceedings, International Conference on Audio-and Video-Based Person Authentication. Citeseer, 1999, pp. 176–181.

[20] B. Clarkson, N. Sawhney, and A. Pentland, "Auditory context awareness via wearable computing," Energy, vol. 400, no. 600, p. 20, 1998.

[21] C. Che and Q. Lin, "Speaker recognition using hmm with experiments on the yoho database," in Fourth European Conference on Speech Communication and Technology, 1995.

[22] K. P. Murphy, "Active learning of causal bayes net structure," 2001.

[23] N. Poh and J. Korczak, "Hybrid biometric person authentication using face and voice features," in International Conference on Audio-and Video-Based Biometric Person Authentication. Springer, 2001, pp. 348–353.

[24] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 815–823.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1. IEEE, 2005, pp. 886–893.

[26] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2892–2900.

[27] E. Verteletskaya and K. Sakhnov, "Voice activity detection for speech enhancement applications," Acta Polytechnica, vol. 50, no. 4, 2010.

[28] S. Kumar, K. Singh, and R. Saxena, "Analysis of dirichlet and generalized "hamming" window functions in the fractional fourier transform domains," signal processing, vol. 91, no. 3, pp. 600–606, 2011.

[29] M. R. Hasan, M. Jamil, M. Rahman et al., "Speaker identification using mel frequency cepstral coefficients," variations, vol. 1, no. 4, 2004.

[30] S. A. Alim and N. K. A. Rashid, "Some commonly used speech feature extraction algorithms," in From Natural to Artificial Intelligence, R. Lopez-Ruiz, Ed. Rijeka: IntechOpen, 2018, ch. 1. [Online]. Available: https://doi.org/10.5772/intechopen.80419

[31] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of computational and applied mathematics, vol. 20, pp. 53–65, 1987.