# Integration of sensory modalities for advice in human-robot scenarios

Francisco Cruz[1,2]

[1] School of Computer Science and Engineering, UNSW, Sydney, Australia.
[2] Escuela de Ingenieria, Universidad Central de Chile, Santiago, Chile.
f.cruz@unsw.edu.au

**Abstract**

Robots in domestic environments are receiving more attention, especially in scenarios where they should interact with parent-like trainers for dynamically acquiring and refining knowledge. In learning approaches, a promising extension has been to incorporate an external parent-like trainer into the learning cycle in order to scaffold and speed up the apprenticeship using advice about what actions should be performed to achieve a goal. Different uni-modal control interfaces have been proposed that are often quite limited and do not take into account multiple sensor modalities. In this paper, we propose the integration of audiovisual patterns to provide advice to the agent using multi-modal information. In our approach, advice can be given using either speech, gestures, or a combination of both. We introduce a mathematical model to integrate multi-modal information from uni-modal modules based on their confidence. Results show that multimodal integration leads to either strengthen or diminish the integrated confidence value in comparison to the uni-modal approaches.

## 1 Motivation and Research Problem

Human-Robot Interaction (HRI) has become an increasingly interesting area of study among developmental roboticists since robot learning can be speeded up with the use of parent-like trainers who deliver useful advice allowing robots to learn a specific task in less time than a robot exploring the environment autonomously [1, 2]. When interacting with parent-like trainers, robots are subject to different environmental stimuli which can be present in various modalities. In general terms, it is possible to think about some of those stimuli as guidance that the parent-like trainer delivers to the apprentice agent [3]. Nevertheless, when more modalities are considered, issues can emerge regarding the interpretation and integration of multi-modal information, especially when multiple sources are conflicting or ambiguous [4]. As a consequence, the advice may not be clear and misunderstood, and hence, may lead the apprentice agent to a decreased performance when solving a task [5] due to poor action selection [6].

People are constantly subject to different perceptual stimuli through different modalities such as vision, hearing, and touch among others. Such modalities are used to perceive information and process it independently, in parallel, or integrating the received information to provide a coherent and robust perceptual experience [7]. Similarly, humanoid robots work with many of these sensory modalities and the way of processing and integrating the information coming from various sources is currently an important research issue in autonomous robotics [8]. In HRI scenarios, robots can take advantage of such multi-sensory information in order to improve their capabilities when any sensory modality is limited, lacking, or unavailable [9]. Nevertheless, in domestic scenarios and dynamic environments, assistive robot companions still need to understand and interpret instructions faster and more efficiently, yielding the integration of available multi-sensory information with different confidence levels in a consistent mode.

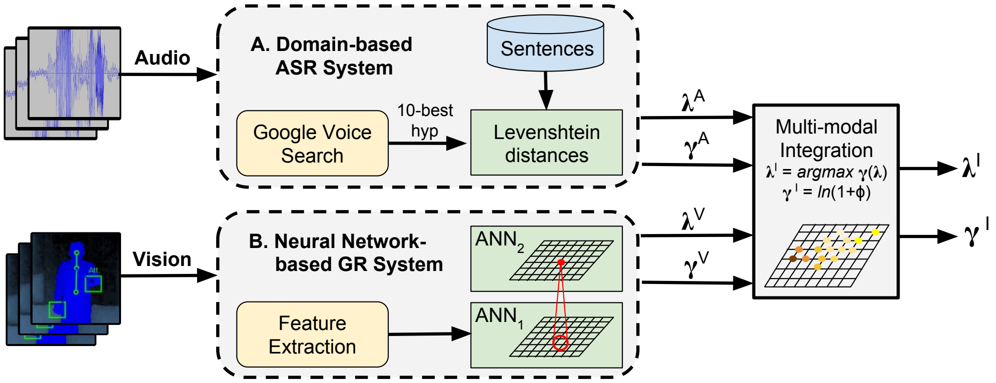

Figure 1: Overall view of the system architecture [10]. A domain-based automatic speech recognition system (on top) processes the audio input modality and a neural network-based gesture recognition system (at bottom) processes the visual input modality. Afterwards, they become the input of the multi-modal integrative system obtaining the integrated advice and confidence value.

## 2 Contribution and Discussion

In our architecture, a parent-like trainer interacts with an apprentice robot using speech and gestures as guidance. Therefore, we are particularly focused on the integration of multi-modal audio-visual inputs. A general overview of the architecture including the speech and gesture processing is depicted in Fig. 1, where $\lambda$ and $\gamma$ are the label and the confidence value respectively. First, the audio and visual sensory inputs are individually processed using the *DOCKS* speech recognition system [11] and a variation of *HandSOM* for gesture recognition [12]. Then, the outputs, i.e. predicted labels and confidence values, become inputs for the multi-modal integration system.

We propose a mathematical model which relates the predicted advice classes and confidence pairs from uni-sensory input denoted as $(\lambda^A, \gamma^A)$ for audio and $(\lambda^V, \gamma^V)$ for vision. The integrated predicted label $\lambda^I$ is calculated according to the highest confidence value as $\lambda^I = \text{argmax}\ \gamma(\lambda)$. In other words, if the audio and visual labels $\lambda^A$ and $\lambda^V$ are different, then the integrated label $\lambda^I$ takes the value from the modality which has the biggest confidence value.

On the other hand, the integrated confidence value is computed by as $\gamma^I = \ln(1+\phi)$, where $\phi$ is a time-varying parameter which depends on each label $\lambda$ and confidence value $\gamma$. We call this parameter the *likeness parameter* and it is obtained according to:

$$\phi = \begin{cases} \gamma^A + \gamma^V & \text{if } \lambda^A = \lambda^V \\ |\gamma^A - \gamma^V| & \text{if } \lambda^A \neq \lambda^V \end{cases} \tag{1}$$

If the labels $\lambda^A$ and $\lambda^V$ are the same, then the confidence value $\gamma^I$ is strengthened. On the contrary, if the labels $\lambda^A$ and $\lambda^V$ are different, then the integrated confidence value $\gamma^I$ is diminished given the differences in the classification.

Therefore, in this work, we have proposed a multi-modal integration of dynamic audiovisual input advice. In this regard, we have shown an integration function that allows to strengthen or diminish the integrated advice for a learning robot using multiple sources of information for a more natural trainer-like learning procedure. The higher (or lower) confidence value of the integrated signal can lead the robot to act differently according to the specific task that it is intended to solve. Future work directions consider experiments in HRI scenarios including online interactions and contextual affordances [13] in order to effectively test the proposed method.

# References

[1] F. Cruz, P. Wüppen, S. Magg, A. Fazrie, and S. Wermter, "Agent-advising approaches in an interactive reinforcement learning scenario," in *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 209–214, IEEE, 2017.

[2] C. C. Millan-Arias, B. J. Fernandes, F. Cruz, R. Dazeley, and S. Fernandes, "A robust approach for continuous interactive actor-critic algorithms," *IEEE Access*, vol. 9, pp. 104242–104260, 2021.

[3] A. Bignold, F. Cruz, R. Dazeley, P. Vamplew, and C. Foale, "Human engagement providing evaluative and informative advice for interactive reinforcement learning," *Neural Computing and Applications*, vol. 35, no. 25, pp. 18215–18230, 2023.

[4] J. Bauer, J. Dávila-Chacón, and S. Wermter, "Modeling development of natural multi-sensory integration using neural self-organisation and probabilistic population codes," *Connection Science*, vol. 27, pp. 358–376, 2015.

[5] A. L. Thomaz and C. Breazeal, "Asymmetric interpretations of positive and negative human feedback for a social learning agent," in *Proceedings of IEEE International Symposium on Robot and Human Interactive Communication RO-MAN*, pp. 720–725, 2007.

[6] A. Ayala, C. Henríquez, and F. Cruz, "Reinforcement learning using continuous states and interactive feedback," in *Proceedings of the 2nd International Conference on Applications of Intelligent Systems*, pp. 1–5, 2019.

[7] M. Andre, V. G. Popescu, A. Shaikh, A. Medl, I. Marsic, C. Kulikowski, and J. Flanagan, "Integration of speech and gesture for multimodal human-computer interaction," in *Proceedings of International Conference on Cooperative Multimodal Communication*, pp. 28–30, 1998.

[8] D. Kimura and O. Hasegawa, "Estimating multimodal attributes for unknown objects," in *Proceedings of International Joint Conference on Neural Networks IJCNN*, pp. 1–8, 2015.

[9] Y. Ozasa, Y. Ariki, M. Nakano, and N. I. Martinetz, "Disambiguation in unknown object detection by integrating image and speech recognition confidences," in *Proceedings of Asian Conference on Computer Vision*, pp. 85–96, 2012.

[10] F. Cruz, G. I. Parisi, J. Twiefel, and S. Wermter, "Multi-modal integration of dynamic audiovisual patterns for an interactive reinforcement learning scenario," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 759–766, IEEE, 2016.

[11] J. Twiefel, T. Baumann, S. Heinrich, and S. Wermter, "Improving domain-independent cloud-based speech recognition with domain-dependent phonetic post-processing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, 2014.

[12] G. I. Parisi, D. Jirak, and S. Wermter, "Handsom-neural clustering of hand motion for gesture recognition in real time," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 981–986, IEEE, 2014.

[13] F. Cruz, G. I. Parisi, and S. Wermter, "Learning contextual affordances with an associative neural architecture.," in *ESANN*, 2016.