Predictive Models for Robot Ego-Noise Learning and Imitation

Antonio Pico Villalpando
Adaptive Systems Group
Humboldt-Universität zu Berlin
Berlin, Germany
pivillaa@informatik.hu-berlin.de

Guido Schillaci
Adaptive Systems Group
Humboldt-Universität zu Berlin
Berlin, Germany
guido.schillaci@informatik.hu-berlin.de

Verena V. Hafner
Adaptive Systems Group
Humboldt-Universität zu Berlin
Berlin, Germany
hafner@informatik.hu-berlin.de

***Abstract*—We investigate predictive models for robot ego-noise learning and imitation. In particular, we present a framework based on internal models—such as forward and inverse models—that allow a robot to learn how its movements sound like, and to communicate actions to perform to other robots through auditory means. We adopt a developmental approach in the learning of such models, where training sensorimotor data is gathered through self-exploration behaviours. In a simulated experiment presented here, a robot generates specific auditory features from an *intended* sequence of actions and communicates them for reproduction to another robot, which consequently decodes them into motor commands, using the knowledge of its own motor system. As to the current state, this paper presents an experiment where a robot reproduces auditory sequences previously generated by itself. The presented experiment demonstrates the potentials of the proposed architecture for robot ego-noise learning and for robot communication and imitation through natural means, such as audition. Future work will include situations where different agents use models that are trained with—and thus are specific to—their own self-generated sensorimotor data.**

I. Introduction

Ego-noise—the auditory noise that an agent produces while moving around—is usually considered as a negative feature in robot audition that needs to be cancelled or removed [1], [2]. Indeed, ego-noise disturbs the auditory input signals captured from the microphones, and can severely decrease the performance of a speech recogniser or of a sound source localiser. However, as we investigated in previous works [3]–[5], robot ego-noise does not only have a negative impact in robot audition. In fact, it can carry out very useful information, such as about the movements that the robot is executing and about characteristics of its surrounding environment.

In a previous study [5], we proposed a biologically inspired model for coding internal body representations that can generate predictions of auditory and motor experience in a humanoid robot, and we demonstrated how such ego-noise predictions can allow the implementation of basic cognitive skills. In fact, we observed that being able to estimate self-induced changes in the auditory signal is crucial both for attenuating ego-noise, and thus for enhancing the auditory signal for further processing such as speech recognition, but also for distinguishing ego-noise from other sounds in natural acoustic environments, which is a prerequisite for efficient and intuitive interaction with other people and with the surrounding [5].

Similarly, in [3], we implemented predictive forward models [6], [7] as computational tools for encoding the dynamics of the motor system of a custom mobile robotic platform and the effect of self-produced movements on the perceived ego-noise. We tested the predictive capabilities of the models in two experimental settings: first, we demonstrated that ego-noise predictions can be used for classifying velocity profiles from auditory signals the robot is listening to; second, we showed how the auditory predictions can be used also to detect changes in the external environment. As an example, we demonstrated how the robot could detect a change in the inclination of the surface where it was moving around.

In the aforementioned studies [3], [5], we represented the ego-noise produced by the robot movements using Mel-frequency Cepstral Coefficients (MFCCs), which are features derived from a type of cepstral representation of the auditory signal commonly used in speech recognition [8], as well as for music classification [9]. In a following work [4], we investigated the adoption of Convolutional Autoencoders (CAEs [10]) for semi-supervised feature learning of the auditory signals[1], as an alternative to MFCCs. Moreover, we proposed an implementation of an inverse model for encoding the mapping between the auditory noise and the motor commands that generated it, in the context of an imitation experiment. We carried out two experiments [4]. In a first imitation test, we recorded the ego-noise of a mobile robotic platform while it was performing a predefined set of command sequences. After having trained the inverse models, the robot was asked to reproduce the movements it previously performed by inferring the motor commands from the auditory information. During the prediction phase, the inverse model was fed with auditory features generated through a Convolutional Autoencoder and with the current robot sensory state. In a second test, the robot imitated the movements from listening to a prerecorded noise reproduced by a loudspeaker.

It has to be noted that the models presented in [3]–[5] were all trained with data generated through a self-exploration

[1]We implemented the CAE using Keras (https://keras.io/) and Theano (https://github.com/Theano/Theano) libraries.

behaviour, namely motor babbling, which is inspired by a behaviour exhibited by infants during early developmental stages. Through self-exploration, infants gather knowledge about their body capabilities and acquire coordination skills [11], [12]. Self-exploration is an essential behavioural component for the acquisition of sensorimotor experience in robots as well, and thus for the autonomous formation of internal models and body representations [7].

This work represents a follow-up of the studies described above [3]–[5]. Similarly, here we investigate the learning of predictive internal models—of both inverse and forward models, as in the classic framework for motor control [6]—in a custom robotic platform (see Figure 1). Following the same developmental approach, we equip the robot with a simple self-exploration behaviour, namely random babbling[2], for the autonomous generation of auditory and motor training samples.

These training samples are used here to train four models: (1) a Convolutional Autoencoder, for semi-supervised feature learning of the auditory signals; (2) an inverse model, for encoding the controller of the robot, which maps initial sensory states (wheels speed, wheels differential speed and last executed motor commands) and auditory features with the motor command that produced that specific auditory signal; (3-4) two forward models, each mapping the initial sensory state of the robot and the applied motor command onto two resulting sensory outcomes (3): the produced auditory signal, encoded with the CAE; (4): the resulting velocities of the robot wheels.

In [4], a CAE was adopted to compress auditory signals into a feature vector, and for inferring through an inverse model the motor commands that generated that specific ego-noise. It was argued that such an approach could allow for basic imitation behaviours, where a robot listening to a specific ego-noise could reproduce the same action (if produced by an agent with similar embodiment). A similar study on autoencoders and sensorimotor learning was presented in [13], although on a classical navigation task. Winfield and Erbas [14] discussed robot-robot imitation of movements, sounds and lights. Here, we extend the previous work by investigating the generative capabilities of the adopted models. In fact, the joint action of the forward model and the Convolutional Autoencoder allows the robot to *generate* a specific ego-noise *without* actually performing the action. This would, for instance, enable a robot to *communicate* to another robot a determined sequence of actions to perform, through auditory means. In the experiment presented here, we simulate the condition where a robot A communicates through auditory means an intended sequence of actions to a robot B, which subsequently decodes the received ego-noise into specific motor commands through its own inverse model (see figure 2). It can be argued that motor commands can be transmitted in robot through other means, such as wireless communication. However, the scope of this work is not to find the best solution for robot communication *per se*, rather to show that brain inspired computational models can serve as a prerequisite for communication and imitation, in the particular complex problem of auditory communication. Sound transmission for imitation purposes can seem quirky and prone to dangerous interferences, especially in presence of small mechanical differences in the robots' motors. Humans are highly performant in communication and imitation processes using auditory means (i.e. musicians' skill in coordinated play), even when the embodiment's characteristics between individuals or between the tools they are using (e.g. music instruments) are highly different. This work aims at providing insights in the understanding of auditory imitation processes, and it focuses on ego-noise as a sample source of auditory information. Moreover, we strongly believe that studying sensorimotor development in the context of robot audition and ego-noise is very relevant for the ICDL-EpiRob community, ego-noise being a result of the particular embodiment of a robot and of the environment it interacts with. Currently, as it will be discussed later in the paper, this capability has been tested only with forward and inverse models trained on the same sensorimotor dataset gathered from a single robot. Thus, the imitation experiment is just a simulation of a communication between two robots, which here are simply the same robots with the same internal models. We are currently carrying out an experiment, not yet reported here, where each of the robots utilises its own forward and inverse models and CAE, that is where it uses models that are trained with—and thus are specific to—their own self-generated sensorimotor data. Nonetheless, the presented experiment already demonstrates the potentials of the proposed architecture for robot ego-noise learning and for robot communication and imitation through natural means, such as audition.

The rest of the paper is structured as follows. In section II, we introduce the methodology behind this work. We then describe the experiment in section III and present the results of this work in section IV. Finally, we conclude the work in section V where we depict our future steps in this research.

II. Methodology

This section describes the methodology adopted in this work.

A. Robotic platform

We used a wheeled robot [4] developed at the Adaptive Systems Group at the Computer Science Department of the Humboldt-Universität zu Berlin (Figure 1).

The robot is equipped with two DC gear motors placed in a differential configuration, as well as a microphone for audio recording. Their characteristics are slightly different from the ones of the platform developed in [4]. In particular, we adopted here more resistant and powerful motors (Pololu Metal Gearmotor 25Dx52L) and a better microphone (RODE SmartLav+).

[2]More efficient exploration strategies have been proposed in the literature (see [7] for a review).

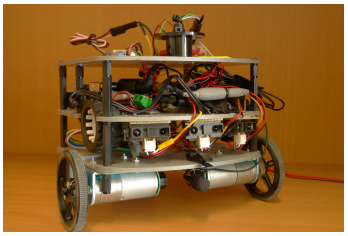

Fig. 1. The wheeled robot developed for the experiment.

Each motor has attached a quadrature magnetic encoder in order to provide a speed feedback to the system (we used the number of encoder counts made in a range of 100ms as a velocity measure). ROS (Robot Operating System) has been used for software development.

B. Ego-noise representation

As in [4], we implemented a semi-supervised learning process to represent auditory features, by means of a Convolutional Autoencoder [10].

An autoencoder is a neural network that tries to reconstruct its input [15], and is composed of two components: an encoder, which compresses the input into a feature vector of lower dimensionality, and a decoder, which reconstructs the input from the compressed vector. Learning processes in autoencoders are designed in a way that they generate features that store only useful properties of the data [4], [15].

C. Model Architecture

Figure 2 depicts the model architecture behind the experiment presented in the following section. In particular, the architecture is mainly composed of four models:

(1) A convolutional autoencoder adopted for compressing auditory signals into auditory features, and for decoding them back into auditory signals.
(2) An inverse model for mapping the initial sensory state of the robot (wheels velocity) and a feature vector encoding an auditory chunk, into the motor command that generated such an ego-noise.
(3) A forward model mapping the initial sensory state of the robot and the motor commands applied to the wheels, into the resulting auditory features.
(4) A forward model mapping the initial sensory state of the robot and the motor commands applied to its wheels, into the resulting new velocity of the wheels.

The forward model (4) has been implemented in order to allow long-term predictions. In fact, as depicted in figure 2, the output of the forward model (4) is fed back to the forward model (3) at the following time step. This allows the execution of further auditory predictions, and thus the generation of sounds from intended *sequences* of movements. The forward and inverse models are implemented as Multilayer Perceptrons.

Figure 2 shows how the framework can be used for an imitation experiment. In particular, we investigate the generative capabilities of the adopted models. In fact, the joint action of the forward model and the convolutional autoencoder allows the robot to *generate* a specific ego-noise (or a sequence of ego-noise chunks, where wheel speed predictions are fed back in a loop into the same forward model) *without* actually performing the action.

Figure 2 also illustrates how a robot can *communicate* to another robot a determined sequence of actions to perform, through auditory means. Forward model predictions, in the form of auditory features, can be restored as FFT signals through the decoder of the CAE and eventually re-transformed into auditory signals to be reproduced by a loudspeaker. A second robot could thus listen to the generated noise, transform it back—through its encoder—into an auditory code, and estimate the intended motor commands through its inverse model. In the experiment presented here, we simulate the condition where a robot A communicates through auditory means an intended sequence of actions to a robot B, which subsequently decodes the received ego-noise into specific motor commands through its own inverse model.

Currently, this capability has been tested only with forward and inverse models trained on the same sensorimotor dataset gathered from one single robot. Thus, the imitation experiment is just a simulation of a communication between two robots, which have the same internal models. We are currently carrying out experiments where two different robots use internal models trained with each own self-generated sensorimotor data.

III. Experiment

The experimental setup consists of the wheeled robot depicted in Figure 1 performing movements in a rectangular arena (1.8 m x 1.0 m) made of a flat wooden floor and cardboard walls. In particular, we used the same training procedure adopted in our previous work [4], as described in the following.

We report here the description of the data collection process. During the learning phase, the robot executed five different random babbling behaviours:

1. A random motor command is sampled from a uniform distribution (range $[0, 10]$, i.e. only positive values, that is forward movements) and applied equally to both wheels. The same command is executed every 150 milliseconds for n iterations. n is sampled from a uniform distribution (range $[1, 15]$).
2. Exploration behaviour (1) is applied only to the left wheel. The right wheel is kept on hold (null speed).
3. Exploration behaviour (1) is applied only to the right wheel. The left wheel is kept on hold (null speed).
4. Exploration behaviour (1) is applied to both wheels. At each iteration, a different value is sampled for each motor.
5. Both wheels are stopped, producing silence.

Each behaviour was executed 5000 times, resulting in 25,000 samples.

During the exploration behaviours, auditory signals were recorded with the embedded microphone RODE SmartLav+. The input signal was captured at a frequency rate of 22,050

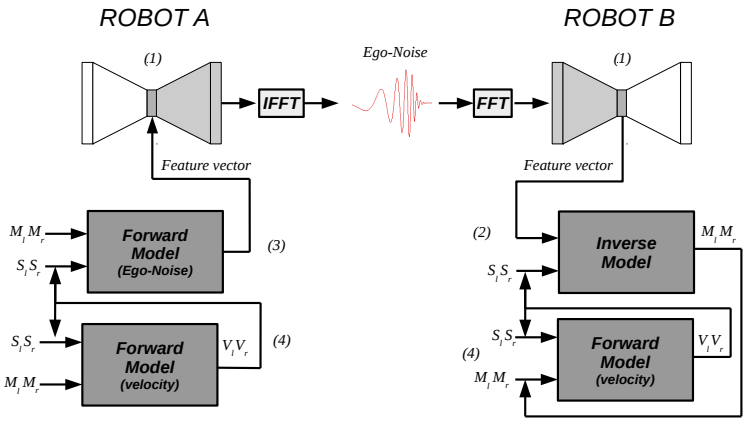

Fig. 2. The forward models architecture (Autoencoder (1), inverse model (2), ego-noise forward model (3), speed forward model (4)), characterised by the following input and output variables. **S**: Sensory state at time {t}, composed of the velocity V at time {t}, the differential of the velocity (between time {t} and time {t-1}) and the motor command applied at time {t-1}. A sensory state is read for both wheels: left (l) and right (r). **M**: Motor command applied at time {t} to both wheels: left (l) and right (r). **V**: Velocity of the left wheel (l) and the right wheel (r). When V is the output of the forward model, than it is related to the time step {t+1}. **Ego-Noise (FFT)**: a Fast Fourier Transform (FFT) operator has been applied to each 2048 samples auditory chunk, resulting in a vector of 1025 values. **Feature vector**: The compressed vector from the latent space of the Conditional Autoencoder

Hz. In particular, we recorded 2048 samples every 150 milliseconds. After every recording there was a pause of 50ms. This time is needed for the robot to calculate the output of the models when is used in real time. We adopted the same auditory, motor and proprioceptive synchronisation framework developed in [3] and [4]. A Fast Fourier Transform (FFT) operator has been applied to each 2048 samples auditory chunk, resulting in a vector of 1025 values.

The gathered data was then used to train a Convolutional Autoencoder characterised by a latent space dimensionality of 5. This means that the CAE is capable of compressing a chunk of 1025-dimensional FFT vector into a vector of 5-dimensions.

The 25000 samples dataset was also used for training the inverse and forward models. Further 4500 samples, generated through the aforementioned exploration behaviours (randomly chosen), were collected for the testing phase.

The tests consisted in the following procedure. We want Robot A to communicate an action to perform to Robot B through auditory means.

In order to generate the auditory data to be sent to robot B, we took the motor command sequences from the recorded test samples, and used them as the intended motor sequence to transmit. This command sequence was then fed into the robot A forward models (see Figure 2) to infer the new motor speed and auditory states. In the initial first step, the values of the current speed, the differential speed and the previous motor command of both motors were set to zero. These data, together with the motor command taken from the test sequence, were fed into the Robot A ego-noise (3) and velocity (4) forward models (see figure 2). The output of the models is the ego-noise feature vector and speed of the motors that would result if the input motor commands were executed. The simulated sensory state (speed, differential of speed) along with the executed motor command were then fed back into the Robot A forward models, as well as the next motor command that would be executed in the test sequence. This process was repeated until the 4500 motor commands were simulated. Through this process, we generated a series of 4500 auditory features. In the next step, we sent them to robot B.

The first step for communicating the information to robot B consisted in decoding the auditory feature vector, by using the decoder part of the pre-trained CAE. We thus obtained a series of 1025-dimensional chunks consisting of the magnitude of the FFT chunks. To generate the ego-noise signals in the time domain, we applied an inverse FFT operator, using random numbers from $-\pi$ to π as the phase vector. In order to simulate loss of fidelity in the transmission of the signal, we added random gaussian noise with a standard deviation of 0.1 times the standard deviation of the original signal.

The receptor robot (robot B in Figure 2) took the audio signal—in the experiment presented here, the simulated signal was simply passed forward, instead of being reproduced by a loudspeaker—, applied an FFT operator and used the autoencoder to get the auditory feature vector. The vector was fed into the inverse model (starting with initial sensory conditions, that is speed, differential and previous motor command, as set to zero), to infer the motor commands that were needed to generate such an input ego-noise. To generate the following sensory states, the motor commands were fed into the velocity (4) forward model. The sensory output was then fed back into the same forward model and also into the inverse model, along with the next ego-noise features. This process was repeated until we generated the complete motor command sequence we wanted Robot B to imitate.

IV. Results

Figures 3, 4, 5, 6 and 7 illustrate the results of this study. In particular, figure 5 shows the velocity forward models predictions for the robot A and robot B (so after the full simulation process). Figures 3 and 4 show the mean squared

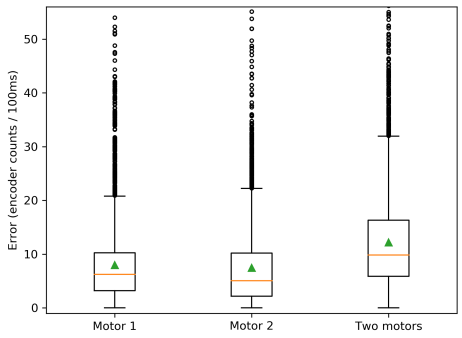

Fig. 3. Mean Squared Errors of robot A velocity (3) forward model for motor 1, 2 and both motors.

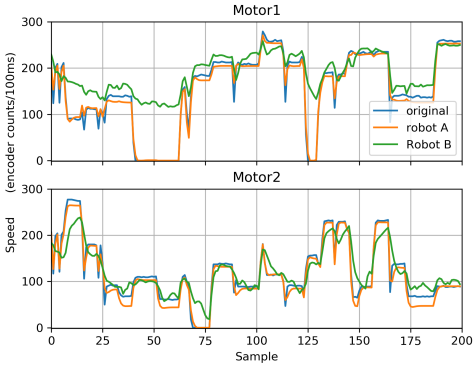

Fig. 5. Comparison between the original velocity signal (blue) and the signals generated by robot A (orange) and B (green) velocity forward models. Last 200 elements of the sequence are shown.

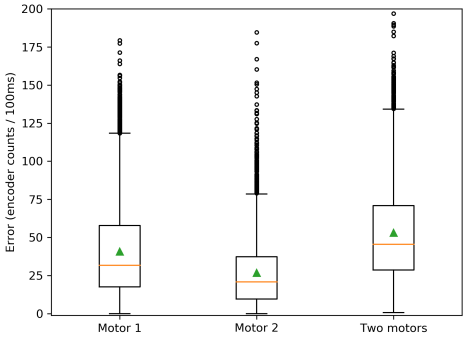

Fig. 4. Mean Squared Errors of robot B velocity (3) forward model for motor 1, 2 and both motors.

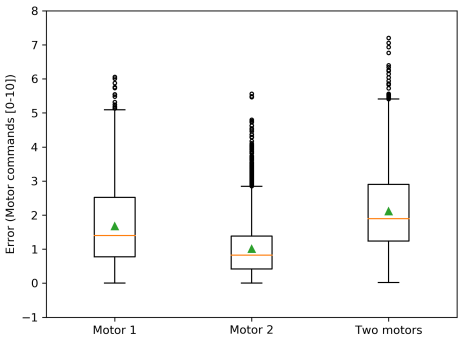

Fig. 6. Mean Squared Errors of the imitated motor commands for motor 1, 2 and both motors (robot B).

errors of these predictions in comparison with the original signal taken from the test dataset.

Figure 5 illustrate the predictions of the velocity (4) forward model. The original velocity is plotted for each motor (blue line), together with the velocity prediction of the forward model of the robot A (orange line) and the velocity prediction of the forward model of the robot B (green line). As expected, predictions of robot B were worse than those of robot A, probably due to the errors accumulated through the propagation of the internal simulation and the communication of that from the robot A to the robot B.

It has to be noted that the prediction error did not accumulate in robot A predictions. It could suggest that the last motor command input had a higher importance, than that of the current speed, for the forward model prediction (velocity FM). However, further tests need to be done to confirm this.

Similarly, prediction error did not accumulate in robot B predictions, as well, even though also motor commands were fed back into the loop. This would suggest that action effects did not depend very strongly on the previous sensory states. and that combinations of last speed and last command as inputs helped each other to compensate the error.

As shown in 6, motor predictions were in general better for one of the two motors of the robot, namely motor 2 (right motor), than for motor 1 (left motor). This suggests that the ego-noise produced by one motor, in this case the right one, dominates. A similar result has been obtained in our previous study [4], where one of the two motors dominated.

Figure 7 is perhaps the most important one, as it shows the result of the entire simulation. For each motor, the blue line shows the original motor command sequence taken from the test dataset (figure shows the last 200 samples subset). The orange line depicts the result of the prediction of the inverse model of the robot B, after the entire simulation loop. As visible here, again, predictions were much closer to the original data for the motor 2 (right).

Finally, figure 8 shows the spectrogram of the auditory data produced, respectively, by the sequence, by the forward model of the robot A, and by the forward model of robot B. In fact, we used an ego-noise (3) forward model also to predict what sound the motors of robot B would have done, in order to compare them to the original sound and to the one predicted by robot A

It can be noted, from Figure 8, that the spectrogram of the auditory prediction generated by robot A is pretty similar to

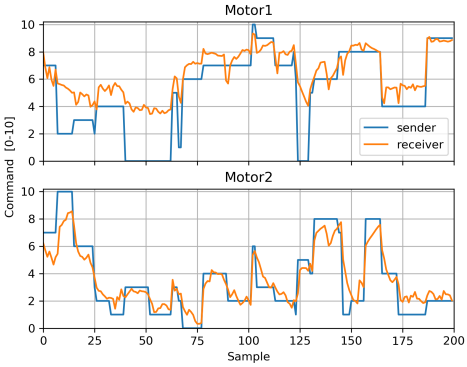

Fig. 7. Comparison between the original motor command sequence (robot A) and the imitated motor command sequence (robot B). Last 200 elements of the sequence are shown.

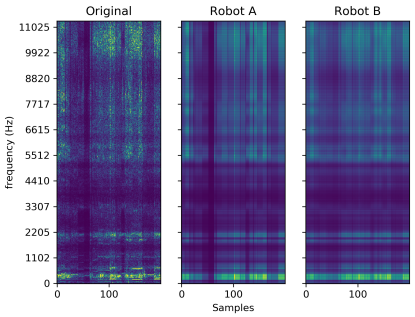

Fig. 8. Ego-noise spectrograms of the original prerecorded sequence (left), sequence generated by robot A forward model (center) and sequence generated by the imitated motor commands of robot B (right). Last 200 elements of the sequence are shown.

the original. Although the one generated by robot B is more diffused, a similar pattern is still visible. Further studies will confirm this similarity in a quantitative way. This analysis has not been carried out so far, as the scope of this work is not to demonstrate the quality of *regenerated* auditory pattern from transmitted auditory codes between robots.

V. Conclusions

We presented a framework for learning and for generating robot ego-noise. A developmental approach was adopted for allowing a wheeled robotic platform to learn the auditory consequences of its own movements. We demonstrated how predictive processes can be used to communicate motor information between robotic agents, through auditory means.

In a simulated experiment presented here, a robot generated a specific auditory feature vector from an *intended* sequence of actions and communicated it for reproduction to another robot, which consequently decoded it into motor commands, using the knowledge of its own motor system. However, this paper presented a preliminary simulation test where a robot reproduces auditory sequences previously generated by itself. We are currently carrying out tests, not yet reported here, where different agents use models that are trained with—and thus are specific to—their own self-generated sensorimotor data. Nonetheless, the presented experiment already demonstrates the potentialities of the proposed architecture for robot ego-noise learning and for robot communication and imitation through natural means, such as audition.

Acknowledgment

This work has partially received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 773875 (EU-H2020 ROMI, Robotics for Microfarms). The models proposed here will be adopted as a basis for sensorimotor learning experiments in the context of microfarming robots.

References

[1] G. Ince, K. Nakadai, T. Rodemann, H. Tsujino, and J.-i. Imura, "A robust speech recognition system against the ego noise of a robot," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[2] H. W. Löllmann, H. Barfuss, A. Deleforge, S. Meier, and W. Kellermann, "Challenges in acoustic signal enhancement for human-robot communication," in *Speech Communication; 11. ITG Symposium; Proceedings of*. VDE, 2014, pp. 1–4.

[3] A. Pico, G. Schillaci, V. V. Hafner, and B. Lara, "How do I sound like? forward models for robot ego-noise prediction," in *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2016, pp. 246–251.

[4] ——, "On robots imitating movements through motor noise prediction," in *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, Sept 2017, pp. 318–323.

[5] G. Schillaci, C. N. Ritter, V. V. Hafner, and B. Lara, "Body representations for robot ego-noise modelling and prediction. towards the development of a sense of agency in artificial agents," in *International Conference on the Simulation and Synthesis of Living Systems (ALife XV)*, 2016, pp. 390–397.

[6] D. M. Wolpert, Z. Ghahramani, and J. R. Flanagan, "Perspective and problems in motor learning," *Trends in Cognitive Sciences*, vol. 5, no. 11, pp. 487–494, 2001.

[7] G. Schillaci, V. V. Hafner, and B. Lara, "Exploration behaviours, body representations and simulations processes for the development of cognition in artificial agents." *Frontiers in Robotics and AI*, vol. 3, no. 39, 2016.

[8] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in {MFCC} computation for speaker recognition," *Speech Communication*, vol. 54, no. 4, pp. 543 – 565, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639311001622

[9] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *JOURNAL OF COMPUTING*, vol. 2, 2010.

[10] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International Conference on Artificial Neural Networks*. Springer, 2011, pp. 52–59.

[11] P. Rochat, "Self-perception and action in infancy," *Experimental brain research*, vol. 123, no. 1-2, pp. 102–109, 1998.

[12] S. Zoia, L. Blason, G. D'Ottavio, M. Bulgheroni, E. Pezzetta, A. Scabar, and U. Castiello, "Evidence of early development of action planning in the human foetus: a kinematic study," *Experimental Brain Research*, vol. 176, no. 2, pp. 217–226, 2007.

[13] J. Olier, E. Barakova, G. Rauterberg, and C. Regazzoni, "Grounded representations through deep variational inference and dynamic programming," in *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2017, pp. 277–282.

[14] A. F. T. Winfield and M. D. Erbas, "On embodied memetic evolution and the emergence of behavioural traditions in robots," *Memetic Computing*, vol. 3, no. 4, pp. 261–270, Dec 2011. [Online]. Available: https://doi.org/10.1007/s12293-011-0063-x

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.