On Robots Imitating Movements Through Motor Noise Prediction

Antonio Pico, Guido Schillaci, Verena V. Hafner
Adaptive Systems Group
Humboldt-Universität zu Berlin
Berlin, Germany
{pivillaa, guido.schillaci, hafner}@informatik.hu-berlin.de

Bruno Lara
Cognitive Robotics Group
Universidad Autónoma del Estado de Morelos
Cuernavaca, Mexico
bruno.lara@uaem.mx

***Abstract*—Robot ego-noise, that is the noise produced while the robot is moving around, can carry useful information about the motor system and the embodiment of the agent. We present an experiment where a mobile robot acquires knowledge about its ego-noise. In particular, we adopt a learning strategy based on self-exploration behaviours and on an inverse model for encoding the mappings between ego-noise and the motor commands that produced it. A Convolutional Autoencoder is adopted for semi-supervised learning of auditory features. The inverse model maps both auditory features and perception of the robot speed to the motor commands that produced the ego-noise. We demonstrate how the trained models can be used for imitating movements from listening to the noise they produce.**

I. Introduction

The capability to imitate the actions of others is an important tool in social learning in humans, as it can dramatically speed up the acquisition of new skills. Scientists are still debating on when and how infants begin to match the behaviours of others [1] and on what is innate or developed in this capability [2][3]. Studies suggest the involvement of the human motor system in the understanding and imitation of actions [4]. The main proposal, partially supported by the discovery of mirror neurons [5], claims that observing and imagining an action excites the motor regions in the brain that are used to execute that same action [6].

Social learning and imitation are fundamental topics also in cognitive robotics. In fact, mechanisms for learning by imitation can reduce the efforts required by engineers in the implementation of new skills in artificial agents. On the other hand, robots can be used as tools in the investigation of the developmental stages of imitation [7], [8].

Many studies can be found on this topic in the cognitive robotics literature. Demiris and colleagues [9], for example, investigated how to combine developmental and social learning in robots. They proposed a framework based on internal models for encoding motor actions learned from self-exploration and for recognising and imitating actions performed by others [10].

In this paper, we present an experiment where a wheeled robot learns the mapping between the auditory noise it produces while moving and the motor commands that generated it. The learning framework is based on two behavioural and computational components: a self-exploration behaviour and an internal model, in particular, an inverse model.

In a previous work [11], we represented the ego-noise generated by the robot motors by using Mel-frequency cepstral coefficients (MFCCs), which are features inspired by human auditory perception and widely used for speech recognition as well as for music classification [12]. Here, we instead apply a Convolutional Autoencoder (CAE [13]) for semi-supervised feature learning of the auditory signals.

In the experiments described in the next sections, we present an implementation of an inverse model, which maps the auditory features learned with the CAE (using different configurations) and the current sensory state of the robot to the motor commands that produced the ego-noise. We also present two imitation tests. In the first test, the robot records its ego-noise while performing a predefined set of command sequences. Then, the robot is asked to reproduce the movements it previously performed by inferring the motor commands from the auditory information (using the pre-trained inverse model). When predicting, the inverse model is fed with auditory features extracted from the recorded sound using the Convolutional Autoencoder and with the current sensory state of the robot (motor speed). In the second test, the robot imitates the movements from listening to a prerecorded noise reproduced by a loudspeaker. Finally, we analyse the prediction performance of the inverse model under different configurations and runs.

The rest of the paper is structured as follows. Section II describes the built robotic platform and the characteristics of the inverse model and the convolutional autoencoder. We describe the experimental setup in Section III and present the results in Section IV. Finally, we draw the conclusions and outline future research directions in Section V.

II. Methodology

A. Robotic platform

For this research we used a wheeled robot developed at the Adaptive Systems Group at the Computer Science Department of the Humboldt-Universität zu Berlin (Figure 1). We adopted this solution in order to have as much control as possible over the robot's hardware and software, which was crucial to obtain a precise synchronization between motor commands and

auditory data. The robot is equipped with two DC gear motors placed in a differential configuration, as well as a microphone for audio recording. Each motor has a quadrature magnetic encoder attached in order to provide a speed feedback to the system. Additionally, the robot has three infrared distance sensors, which were necessary for avoiding collisions during the learning session that will be described later.

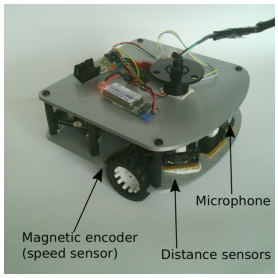

Fig. 1. Mobile robot used in the experiments presented in this paper. The robot is a modified version of the platform presented in [11].

B. The computational framework

We adopted the theoretical framework of internal models for encoding the robot knowledge about its ego-noise and its motor system. As demonstrated by many studies in the literature [14], such models can provide artificial agents with the capability to anticipate the sensory outcomes of intended actions (in the case of forward models) or to infer the specific motor command that produces a desired perceptual experience (in the case of inverse models). Forward and inverse models are well known tools in control theory, where they are often referred to as *predictors* and *controllers*. Recently, they gained interest also in the neuroscience and cognitive robotics communities for the role they can have in explaining brain functionalities and in implementing basic motor and cognitive skills in artificial agents [14].

In a previous paper [11], we presented a mechanism for learning and predicting the auditory consequences of self-generated movements on a custom robotic platform using forward models. We demonstrated how the predictive capabilities of such models can be used to classify motor behaviours based on the perceived auditory signals and to detect unexpected changes in environmental conditions based on simulating the production of ego-noise.

In this work, we use an inverse model for encoding the mapping between the auditory noise and the motor commands that generated it. In particular, the inverse model is implemented as a neural network (see details in Section III). Its input is formed by auditory signals and the speed of the wheels and the output is the motor commands that generated those inputs.

The model is trained on data collected using an exploration behaviour where the robot moves with random motor commands. This behaviour is inspired by the way human infants acquire motor skills during early developmental stages (for a review on exploration behaviours in artificial agents and in humans, please refer to [14]).

C. Ego-noise representation

In a previous work on modelling robot ego-noise [11], we used Mel-Frequency Cepstral Coefficients (MFCCs) for representing the auditory noise generated by the robot motors. While these coefficients yielded good results, they are pre-designed features that require a specific extraction process dependent on a set of parameters.

Here, we investigate how an artificial agent can learn these features in a semi-supervised fashion, by means of a Convolutional Autoencoder [13]. An autoencoder is a neural network that is trained to copy its input to its output [15]. In other words, it is a network that tries to reconstruct its input. Usually, autoencoders are composed by an encoder - i.e. a network that compresses the input into a code of, usually, lower dimensionality - and a decoder - i.e. a network that reconstruct the input from the code. By forcing the network to be unable to copy the input perfectly to the output, the learning process generates codes, or features, that store only useful properties of the data.

An autoencoder with linear activation functions learns to span the same subspace as PCA. Adding nonlinear activation functions makes the autoencoder learn more powerful nonlinear generalizations of PCA [15].

Convolutional autoencoders [13] are a variation of standard autoencoders which use, instead of fully-connected layers, a combination of convolutional and deconvolutional layers. Convolutional neural networks are usually applied in image and visual tasks. Nonetheless, convolutional networks and CAEs have been proven to be powerful tools also in audio classification tasks [16] . In this work, we adopt CAE for auditory features extraction and dimensionality reduction.

III. Experimental setup

The experiment consists of two sessions: a learning session and an imitation session. In the first phase, the robot gathers sensorimotor experience by executing a random motor babbling behaviour. This random behaviour generates auditory, proprioceptive and motor data which is used for training two neural networks: a convolutional autoencoder, for semi-supervised auditory feature learning; and an inverse model, implemented as a Multi-Layer Perceptron, which maps the auditory features learned with the CAE (using different configurations) and the current sensory state of the robot to the motor commands that produced the ego-noise.

In the second session, we perform two imitation tests. In the first test, the robot records its ego-noise while it is performing specific command sequences. Then, the robot is asked to reproduce the previously performed sequences by inferring the motor commands from the auditory information using the trained inverse model. When predicting, the inverse model is fed with auditory features extracted from the recorded sound using the CAE and with the current sensory state of the robot. In the second test, the robot imitates the command sequences by listening to the ego-noise reproduced by a loudspeaker.

The experimental setup consisted in the wheeled robot depicted in Figure 1 performing movements in a rectangular

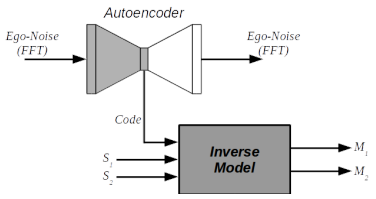

Fig. 2. The connection between the CAE and the IM and their inputs and outputs.

arena (1.8 m x 1.0 m) made of a flat wooden floor and cardboard walls.

In the learning phase, the robot executed five different random babbling behaviours:

1. A random command is sampled from a uniform distribution (range $[0, 10]$, i.e. only forward movements) and applied equally to both motors; a command is sampled every 150 milliseconds for n iterations, where n is sampled from a uniform distribution (range $[1, 15]$).
2. The babbling behaviour (1) is applied only to the left motor. The speed of the right motor is kept to 0.
3. The babbling behaviour (1) is applied only to the right motor. The speed of the left motor is kept to 0.
4. The babbling behaviour (1) is applied to both motors. At each iteration, a different value is sampled for each motor.
5. Both motors are turned off, producing silence.

Each babbling behaviour was executed for the same amount of time, and produced in total 25,000 samples (5,000 per behaviour).

During the babbling behaviours, auditory input signals were recorded with an embedded microphone. The sound was captured at a frequency rate of 22,050 Hz. In particular, 2048 samples were collected every 150 milliseconds, where recording was active for the first 100ms and was held off for 50ms . Particular efforts have been spent on synchronising auditory, motor and proprioceptive signals. For more details, please refer to [11]. A Fast Fourier Transform (FFT) operator has been applied to each 2048 samples auditory chunk, resulting in a vector of 1025 values.

Once having gathered the training data, we trained three different Convolutional Autoencoders, each characterised by a different dimensionality of the latent space: 5, 10 and 20. In other words, CAEs learned auditory features of 5, 10 or 20 dimensions from chunks of FFT vectors with 1025 dimensions.

These features were used as input to three different inverse models, implemented as Multi-Layer Perceptrons. In particular, each inverse model was characterised by the following inputs and outputs:

- Inputs:

$S_1(t)$ Consists of three elements: Left motor speed and acceleration at time t; command sent to the left motor at $t-1$.
$S_2(t)$ Consists of three elements: Right motor speed and acceleration at time t; command sent to the right motor at $t-1$.
$A(t)$ CAE features extracted from the 100ms chunk recorded from time t to time $t + 100ms$.

- Outputs:

$M_1(t)$ Command sent to the left motor at time t.
$M_2(t)$ Command sent to the right motor at time t.

We trained three inverse models: $IM1$, $IM2$ and $IM3$ which are fed with auditory features $A(t)$ of 5, 10 and 20 dimensions, respectively. Figure 2 illustrates the connection between the Convolutional Autoencoder and the inverse model.

For each experiment, three different sequences have been tested, as depicted in Figure 3.

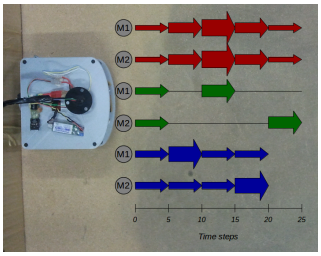

Fig. 3. The three test sequences. In the first sequence, both motors are activated with the same changing speed (the width of the arrow illustrates the magnitude of the velocity). In the second sequence, the following commands are sent: both motors start with the same speed; both stop (silence); only M1 is activated; both motors stop; only left motor is activated. In the third sequence: both motors start with the same speed; left motor increases the speed; both motors go back to the initial speed; right motor increases the speed.

A. Experiment 1

In this experiment, the robot has to repeat a previously executed command sequence based only on the recorded ego-noise produced by the the motors in a relatively quiet environment, using the three inverse models.

The robot moves according to one of the three scripted sequences as shown in Figure 3; while it moves, it records the produced ego-noise. Once having completed the movement, the imitation phase starts. The imitation procedure consists of reading a chunk of auditory data, calculating the FFT and then getting the latent variables from the autoencoder. The extracted features, together with the current sensory state (current speed, acceleration and last command applied to two motors) are fed into the inverse model. A prediction on the inverse model is performed, which outputs the motor commands to be applied in order to produce the ego-noise represented by the latent variables. The process is repeated for each recorded audio chunk. The ego-noise produced by the imitated movements is also logged. We executed 8 runs for each of the three sequences, for a total of 24 runs.

B. Experiment 2

In this second test, the robot has to repeat the same motor sequences as in the previous experiment, but this time a pre-recorded ego-noise is played from a loudspeaker. The speaker was located in front of the robot at ca. 63 cm distance. We performed this test to check whether the imitation performance decreases when the ego-noise is produced by a different source.

The audio track played from the speaker is shorter than the audio from experiment 1, although the same sequence of commands is executed. As in the previous experiment, we

executed 8 runs for each of the three sequences, for a total of 24 runs, with the same procedure described above.

IV. Results

A. Experiment 1

Figure 4 shows the motor commands predicted by each model during the imitation of the three sequences. Each row depicts a different run. The filled areas in the plots show the original commands that had to be imitated. The lines with markers show the motor commands that are inferred by feeding the inverse model with the ego-noise recorded from the robot microphones and the current sensory states perceived by the robot (purple line: IM with 5-D auditory features; orange line: IM with 10-D auditory features; cyan line: IM with 20-D auditory features).

It can be noted that the predictions are similar to the original commands when imitating sequence 1, where the same command is applied to both motors. As for sequence 2 and 3, it can be seen that one of the motors is dominant, in terms of produced noise, compared to the other. For example, in sequence 2, when motor 2 is turned off, the prediction error is quite high for motor 2, probably due to the noise generated by motor 1. After that, when motor 1 is turned off and motor 2 is turned on, the same effect can be observed (prediction error high in motor 1, when it is silent), although with lower magnitude. The same effect can be seen in sequence 3. In spite of this dominance, we found that, at a certain level, the models are capable of distinguish the ego-noise produced by the motor 1 from that produced by the motor 2.

Figure 5 shows the root mean squared error (RMSE) during imitation of the eight runs. It can be seen that the RMSE of motor 2 is in most of the conditions higher than motor 1, suggesting again that motor 1 may have a dominant ego-noise compared to motor 2.

Figure 5 also shows the mean of the RMSEs of the three sequences combined. Here we can see that the performance of the three different models is very similar, being the inverse model with the auditory features of 20 dimensions (IM20) slightly better than the other two models.

B. Experiment 2

Figure 6 shows the motor commands predicted by each model during the imitation of the ego-noise played by the loudspeaker. Playing the ego-noise from loudspeakers decreased the performances of the inverse models in most of the conditions. However, it can be noted that the models were still capable to imitate the command sequences. From the runs of the command sequence 1, it can be seen that the models IM10 and IM20 were not able to predict the highest speed commands, being the model IM05 the one that showed a slightly better performance. It is worth noting that the effect of the dominance of motor 1 was still present in the imitation runs of the sequence 2. Another important observation is that the motor commands generated by the models IM10 and IM20 in the sequences 1 and 3 were not capable of adopting the shape of the original sequence of commands, which resulted in similar command predictions for motors 1 and 2 in both sequences. Figure 7 shows the root mean squared error (RMSE) during imitation of the eight runs for each sequence as well as the mean error of the three sequences combined. The plots show an increment of the prediction errors with respect to the errors calculated in the experiment 1. The mean RMSE of the three different models is also very similar, but in this case the inverse model with auditory features of 5 dimensions (IM05) had a better performance.

C. Assessing similarity of sounds

In order to compare the different sounds used in our experiments, we obtained the Mel-Frequency Cepstral Coefficients (MFCCs) from every recorded audio data and applied the t-sne algorithm for dimensionality reduction and visualisation. Each audio was divided into 18 windows, of which a set of 12 MFCCs was calculated, resulting in a 216 dimensional vector. Figure 8 shows the 2-D visualisation of the features of the sounds generated by the models IM05, IM10 and IM20. The colors of the markers represent the command sequence that generated that sound (red, green and blue for sequence 1, 2 and 3, respectively). From them, four different kinds of audio can be shown:

1) Ego-noise generated by the robot while executing a predefined motor sequence (AR), represented by circle markers 1, 2 and 3.
2) Ego-noise generated by the robot while imitating a previously executed motor sequence (IAR), represented by the x markers 4, 5 and 6.
3) Prerecorded ego-noise (also generated by the robot) played by a loudspeaker (AL), represented by the diamond markers 7, 8 and 9.
4) Ego-noise generated by the robot while imitating the audio played by the loudspeaker (IAL), represented by the $+$ markers 10, 11 and 12.

It can be seen in 8 that the AR audio formed three different clusters, which is consistent with the sequences 1, 2 and 3. The imitation audio IAR also formed three clusters that are close to the respective imitated audio sequences. The AL audio (played by the loudspeakers) appeared on the other side of the plot, far from the audio generated directly by the robot, however the position of the clusters coincided in the vertical dimension with the clusters of the AR audio. It is worth noting that the audio generated by imitating the ego-noise played on loudspeakers (IAL) appeared far from AL but close to the AR audio, forming also three different clusters. Another important observation is that the IAL audio clusters for sequences 1 and 3 of models IM10 and IM20 were not well differentiated from each other, while the IAL clusters of the inverse model IM05 were more separated from each other. This suggests that IM05 had better performance imitating the audio played by the loudspeakers.

V. Conclusion

In this study, we presented a multimodal learning architecture that allows a robot to imitate a previously executed sequence of motor commands by listening to its ego-noise and also to the prerecorded ego-noise played by a loudspeaker.

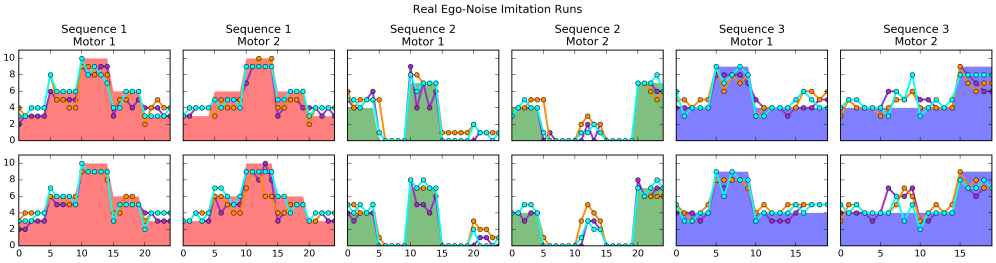

Fig. 4. Plot of two of the eight real ego-noise imitation runs of the command sequences 1, 2 and 3. The filled areas represent the original motor commands. The purple, orange and cyan lines represent the imitation commands executed by motors 1 and 2 of inverse models IM05, IM10 and IM20 respectively.

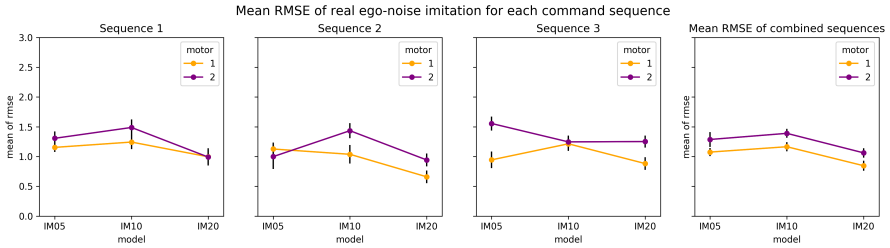

Fig. 5. Mean of the real ego-noise imitation errors (RMSE) obtained from the eight runs of the sequences 1, 2 and 3 with the inverse models IM05, IM10 and IM20. On the right, the mean RMSE of the combined sequences is depicted. The error bars show the 95% confidence interval of the mean.

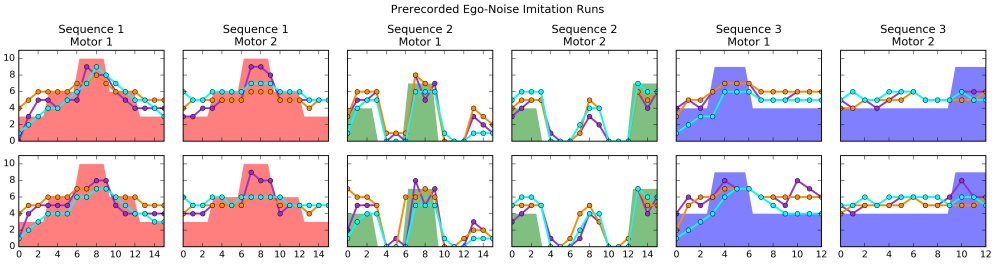

Fig. 6. Plot of two of the eight prerecorded ego-noise imitation runs of the command sequences 1, 2 and 3. The filled areas represent the original motor commands. The purple, orange and cyan lines represent the imitation commands executed by motors 1 and 2 of inverse models 05, 10 and 20 respectively.

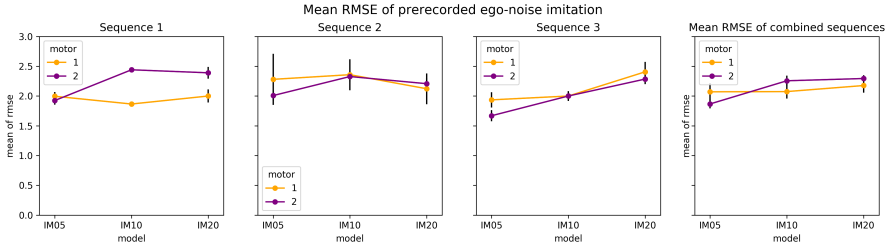

Fig. 7. Mean of the prerecorded ego-noise imitation errors (RMSE) obtained from the eight runs of the sequences 1, 2 and 3 with the inverse models IM05, IM10 and IM20. On the right, the mean RMSE of the combined sequences is depicted. The error bars show the 95% confidence interval of the mean.

In particular, inverse models were adopted as computational tools for encoding robot knowledge about its ego-noise and its motor system. We applied a convolutional autoencoder for a semi-supervised feature learning and dimensionality reduction of the auditory signals. In the experiment, we tested three inverse models with three convolutional autoencoders, each characterised by a different dimensionality of the latent space: 5, 10 and 20. By comparing the root mean squared errors (RMSE) calculated from the imitation of the motor command sequences, we detected a better performance of the model IM20 in the experiment 1 and of the model IM05 in the experiment 2. This was consistent with the analysis results of the audio data generated in the imitation experiments, which indicated that the model IM20 (model with autoencoder code of 20 dimensions) had a better performance when the robot imitated previously executed command sequences (experiment

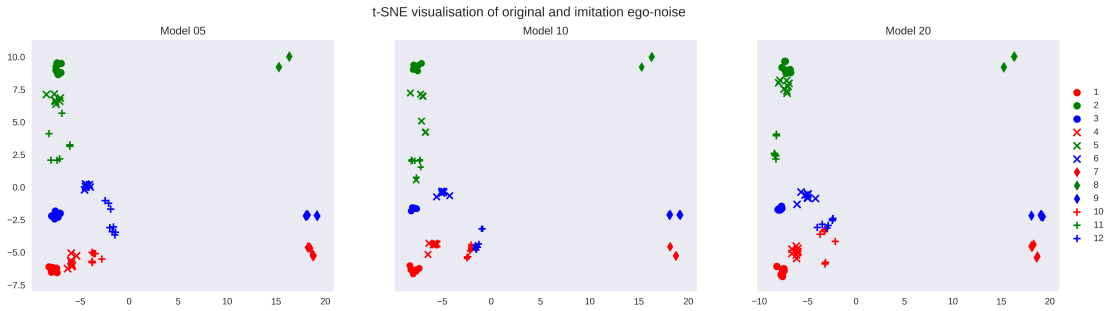

Fig. 8. 2-D visualisation of the original and imitation audio generated from the experiments 1 and 2 with models IM05, IM10 and IM20. Points representing the audio recorded while performing sequence 1 are in red. Points for sequences 2 and 3 are in green and blue respectively. The original audio produced by the robot is represented by the circles (1,2,3). The audio obtained from the imitation of the audio produced by the robot is represented by the x markers (4,5,6). The diamond markers (7,8,9) represent the prerecorded audio played by the loudspeakers. The audio generated from the imitation of that played by the loudspeakers is represented by the $+$ markers (10,11,12)

1), and that the model IM05 (model with autoencoder code of 5 dimensions) performed better when imitating prerecorded ego-noise played by a loudspeaker (experiment 2). This implies that, although a longer audio feature vector generates a better internal model of the robot, it also diminishes the generalization performance.

In the experiments presented in this paper, we trained the CAE and the inverse models separately. We are currently carrying out experiments on training both CAE and IMs in the same run, by connecting the output of the encoder to the input of the inverse model. We expect this solution to outperform the current one, as we expect the feature learned by the CAE to be optimised to reduce the prediction error of the inverse model. On the other hand, we believe that the current solution provides better generalisation capabilities to the model, as the CAE features are not specific to the particular inverse model. In specific implementations, several issues need to be considered. In particular, two motors and wheels are never exactly the same and produce slightly different noise patterns. In our implementation, this allows for the system to learn specific features of the motors and distinguish between them, but might cause problems in an imitation experiment with a different robot. In theory, M_1 and M_2 produce the same sound features, which makes it impossible to predict M_1 and M_2 for different motor velocities. This could be avoided by introducing alternative symmetric motor features $M_3 = |M_1 - M_2|$ and $M_4 = M_1 + M_2$ which contain curvature and speed measures. In this work, we used the encoding section of an autoencoder as an audio feature extractor that fed an inverse model. In further experiments, we will use the decoding section of the autoencoder for converting the audio feature codes (the output of a forward model) into the complete ego-noise frequency spectra. With this, we expect the robot to be able to distinguish its ego-noise from the ego-noise of other robots. This is a feature that can be useful in tasks involving swarm robotics and learning by imitation.

Acknowledgment

The work from Antonio Pico has received funding from the Mexican National Science and Technology Council (CONACyT) and the German Service of Academic Exchange (DAAD).

References

[1] C. Heyes, "Evolution, development and intentional control of imitation," *Philosophical Transactions Of The Royal Society B-Biological Sciences*, vol. 364, pp. 2293–2298, 2009.

[2] S. S. Jones, "The development of imitation in infancy," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1528, pp. 2325–2335, 2009.

[3] A. N. Meltzoff *et al.*, "Imitation of facial and manual gestures by human neonates," *Science*, vol. 198, no. 4312, pp. 75–78, 1977.

[4] A. N. Meltzoff and J. Decety, "What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 358, no. 1431, pp. 491–500, 2003.

[5] G. Rizzolatti, "The mirror neuron system and its function in humans," *Anatomy and embryology*, vol. 210, no. 5-6, pp. 419–421, 2005.

[6] M. Jeannerod, "The representing brain: Neural correlates of motor intention and imagery," *Behavioral and Brain sciences*, vol. 17, no. 02, pp. 187–202, 1994.

[7] A. Cangelosi, M. Schlesinger, and L. B. Smith, *Developmental robotics: From babies to robots*. MIT Press, 2015.

[8] V. V. Hafner and F. Kaplan, "Interpersonal maps: how to map affordances for interaction behaviour," in *Towards affordance-based robot control*. Springer, 2008, pp. 1–15.

[9] M. Johnson and Y. Demiris, "Hierarchies of coupled inverse and forward models for abstraction in robot action planning, recognition and imitation," in *Proceedings of the AISB 2005 Symposium on Imitation in Animals and Artifacts*, 2005, pp. 69–76.

[10] Y. Demiris and A. Dearden, "From motor babbling to hierarchical learning by imitation: a robot developmental pathway," in *Proceedings of the Fifth International Workshop on Epigenetic Robotics*. Lund University Cognitive Studies, 2005.

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

[12] 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.

[13] 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.

[14] G. Schillaci, V. Hafner, and B. Lara, "Exploration behaviors, body representations, and simulation processes for the development of cognition in artificial agents," *Frontiers in Robotics and AI*, vol. 3, p. 39, 2016.

[15] I. Goodfellow *et al.*, *Deep learning*. MIT Press, 2016.

[16] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*. IEEE, 2015, pp. 1–6.