Lab Report: Ego-noise Suppression and Surface Classification in the Aldebaran Nao Robot

Hüseyin Camalan, Dr.Guido Schillaci, Prof.Dr.Verena Hafner

Abstract

Using feedforward neural networks, we implemented a forward model to predict the ego-noise of the walking Aldebaran Nao robot. The predictions of our forward model allow for the subtraction of the ego-noise. In addition to this, using the same predictions, we trained a classifier to distinguish between different walking surfaces. Although this should theoretically be possible, classification performance has been lacklustre. Possible explanations and further attempts to improve the results are outlined below.

Introduction

Figure 1: The Aldebaran Nao Robot.

(Image adapted from https: //www.ald.softbankrobotics.com/sites/aldebaran/files/images/nao_who_is_nao%20%281%29.png)

Ego-noise is the acoustic noise generated by a robot during its movement. It poses a challenge for robotics, because it mixes with the sound from the environment and hinders performance on tasks in which the robot needs to process sound signals (e.g. speech recognition or sound localization). To overcome this problem, it is necessary to be able to predict and "subtract" the ego-noise from the combined sound signal (Schillaci, Ritter, Hafner, & Lara, 2016).

One possible solution for this problem is to use the internal models; specifically, the forward model (Jordan & Rumelhart, 1992). Forward models are used frequently in modeling the motor system because they are able to represent ways in which real life agents can overcome fundamental problems with movement. In a nutshell, forward models allow to predict the future state of a desired variable given the current state of

that variable and of other variable(s) that influence the desired variable. It is thought that real-life agents have such systems that predict their own future states, so that complex tasks can be done without interference from time lags (from sensory processing or motor commands) or noise (for a review, see Miall and Wolpert (1996)).

In our application of the forward model, estimates of the ego-noise are obtained by inputting (i) the angles and (ii) change in angles (i.e. motion) of the robot's joints. Through training the forward model, it is possible to find out how much each joint has an effect on the combined ego-noise.

Using neural networks to implement a forward model, we attempted to predict the ego-noise of the Aldebaran Nao robot (see Figure 1) as it walked on different surfaces. Prediction and subtraction of the ego-noise was already demonstrated in Schillaci et al. (2016) using a single joint (neck joint, this will also be referred to as "head motion"). This project aimed to expand on these previous findings by applying the same procedure in a more complex task, i.e. walking. Additionally, we implemented a classification paradigm to make predictions about the type of surface the robot was walking on.

Materials and Methods

The data collection procedure consisted of placing the robot on one of the specified surfaces (desk, foam tile, carpet) and executing a script, in which the robot took 10 steps in a straight line. Each datapoint, which represented a time step, was temporally placed 20ms apart from one another, and lasted 40ms. Within each datapoint were joint sensor angle, motor command and electric current readings at the beginning and end of the specified period, as well as the corresponding microphone recordings.

The microphone recordings were compressed into *Mel-Frequency Capstrum Coefficient* (MFCC) signals (Davis & Mermelstein, 1980), a widely used compression type in robotics and speech recognition. In a nutshell, MFCC compression sums up frequency values under partially overlapping frequency bands. As a result, a vector much smaller than that sound signal emerges, each one of whose values represent a certain frequency band (this vector is called a filterbank, i.e. an array of filters that represent a signal in a compressed manner). Depending on the spectral organization of the filterbank, the sound signals can be represented in an efficient way without losing much information, which make it computationally easier to analyze these signals. Our MFCC vectors had 13 components, though for space reasons and convenience the last component is omitted in some of the figures.

The sound signal and the joint signals were transmitted from the robot to the computer through different servers, which resulted in a temporal lag. This lag was also present across signals between different trials. Thus, synchronizing signals across signal type and trials was necessary.

All of the code was written in the *python 2.7* programming language, with the help of numerical/scientific computing libraries *numpy* and *scipy* as well as *pybrain*, a library that implements neural networks. Tests were written to ensure that the core functions used for data analysis worked properly.

The data analysis was completed first in the head motion dataset from Schillaci et al. (2011). The purpose of this was to ensure that the scripts for the analysis were functional.

The code used for this project and its results are available in a *git* repository dedicated to the project, accessible at *http://github.com/husejnc/robot_selfsound*.

Results

Synchronization

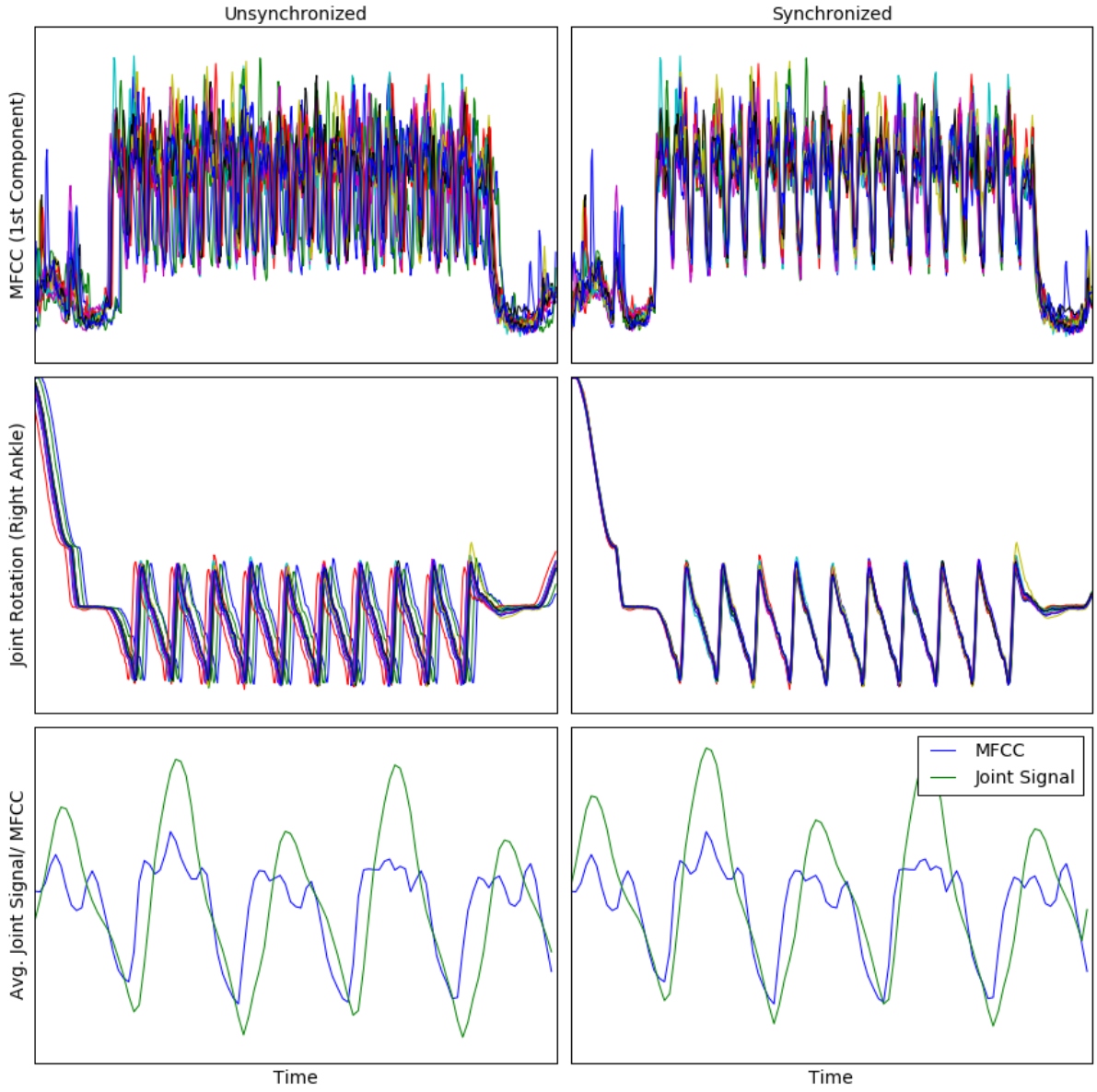

Figure 2: Demonstration of the synchronization process that removes the temporal lags across trials and signal types. Transition from the left to the right column shows the effect of synchronization. Top row: First component of the MFCC signal across all trials. Middle row: Angle of the right ankle (pitch) joint. Bottom row: Comparison of averaged MFCC first component and a "walking signal" in which relevant joint signals are combined.

Results of the synchronization process (described above) is demonstrated in Figure 2. Synchronization was achieved by minimizing absolute squared differences between signals as they were being translated in the time axis. Initially, we attempted to achieve this by removing the time difference between τ=0 and the closest peak of the crosscorrelation function between signal types. The crosscorrelation peaks were not always well-defined,

so this method was not reliable. This synchronization process dealt only with translation, therefore time scales were not altered.

Data Analysis for the Head Motion Experiment (Schillaci et al., 2016)

The mentioned study was able to suppress the ego-noise of a singular joint of the Aldebaran Nao robot using forward models. In addition, a classifier was built successfully on the neural networks that predicted the ego-noise, classifying the head motion speed.

In our study, we initially tested our forward models on this dataset and attempted to replicate the same results to establish a ground truth. To this end, a backpropagating neural network with one intermediate layer was used. The neural network took two inputs: the initial joint angle at the beginning of each timestep and the change in joint angle in the duration of this timestep (obtained by subtracting final angle from the initial angle). With these inputs, an MFCC vector that predicted the sound signal was produced.

Figures 3 and 4 show our prediction results, while Table 1 entails classification performance in a confusion matrix format.

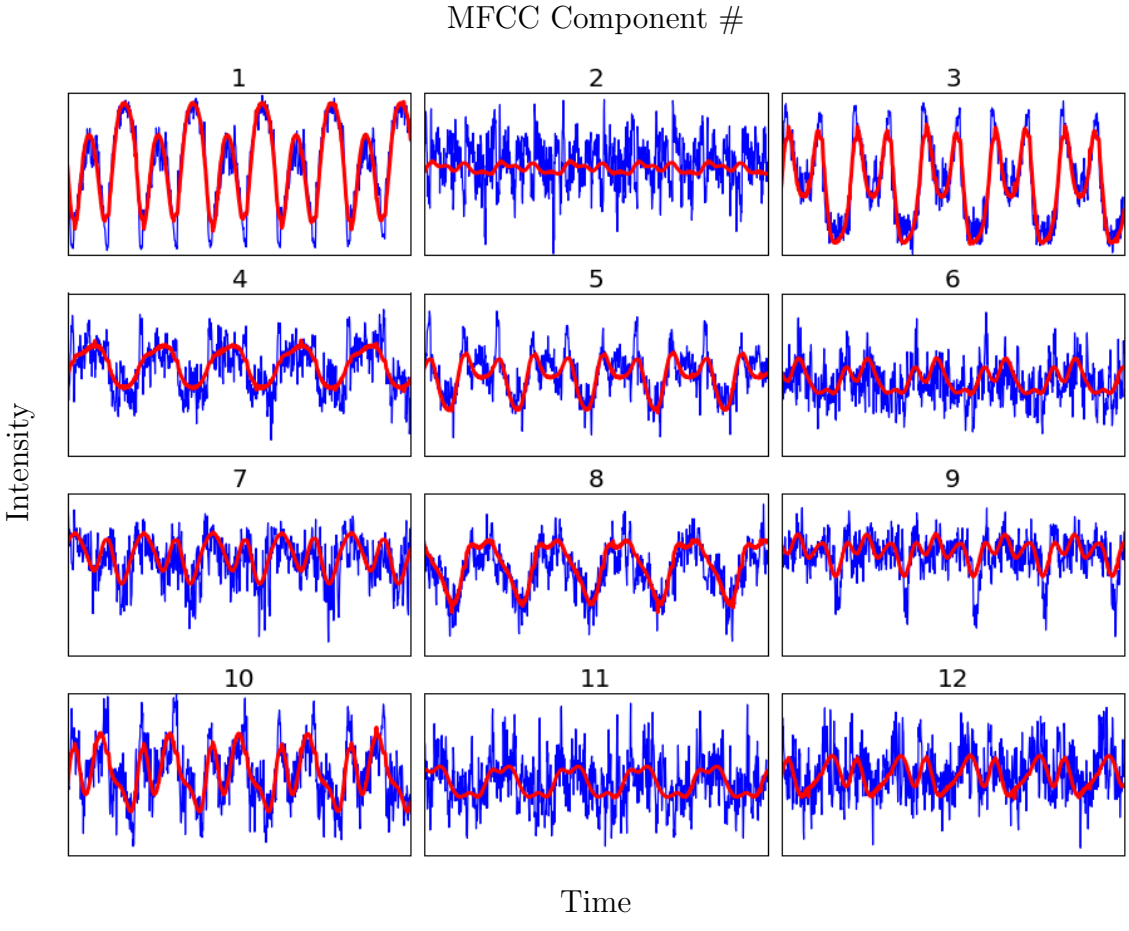

Figure 3: MFCC signals from an arbitrary time segment in a trial (head dataset) and the prediction of the neural network. MFCC components 1-12 are shown for convenience. Blue lines: actual signal. Red lines: predicted signal.

Figure 3 shows a comparison of the real ego-noise and the predictions of the network for a given trial. The training data and testing data were not overlapping. It is important to note that the network is not able to predict the high-frequency fluctuations in the

signal, which are visible in each channel. However, it appears that in some channels this fluctuation is relatively more prominent than others (e.g. second channel vs. first channel).

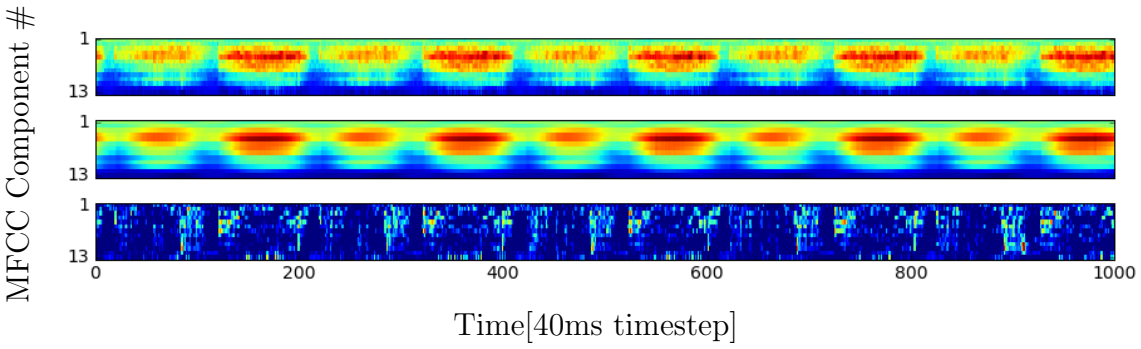

Figure 4: Demonstration of the signal subtraction process for a time segment in the head motion experiment. Up: Heatmap of the original signal. Middle: Heatmap of the predicted signal. Down: Subtracted Signal.

Figure 4 shows the signal subtraction results, where redder colors show higher intensities of sound. Here, MFCC signals are converted into the log-filterbank format, which allows for subtraction. Upon subtraction, spectral flooring is applied to negative log values.

	Slow	Medium	Fast	Very Fast
Slow	.934	.146	.060	.022
Medium	.046	.730	.136	.094
Fast	.008	.022	.744	.126
Very Fast	.010	.100	.060	.756

Table 1: Confusion matrix showing classification performance for predicting head motion speed. Diagonal values show proportion of correct responses for each class. Values are averaged across 5-fold crossvalidation. Inside a crossvalidation fold, data (1000 samples for each class) was divided as %80 for training (800 samples), and %20 for testing (200 samples).

Table 1 shows classification performance from a 5-fold crossvalidation analysis in a confusion matrix format. The signal type here is rotation speed of the head joint. Diagonally placed values indicate the rate of correct predictions. Classification performance ranges between %73-93, in contrast to a chance level of %25.

The classification paradigm is implemented in the following sequence: (i) a separate neural network is trained for each surface category. (ii) The "unobserved" joint data from all surfaces are fed to the networks to create MFCC predictions. (iii) For each data point, the Euclidean distance from the actual MFCC vector to the predictions are computed, and finally the network that produces the least distance is picked as the winner for that point. If the winning network type of the data point is the same as the surface type, it is a correct response, and vice versa.

Walking Experiment

In this experiment we attempted to acquire the same results as above, while facing a more complex problem. First, the ego-noise was formed by multiple joints instead of a single joint (included joints were pitch and roll joints of each of the following for the right and left side: ankle, hip, knee). Second, the classification task consisted of distinguishing between walking surfaces rather than joint motion speeds.

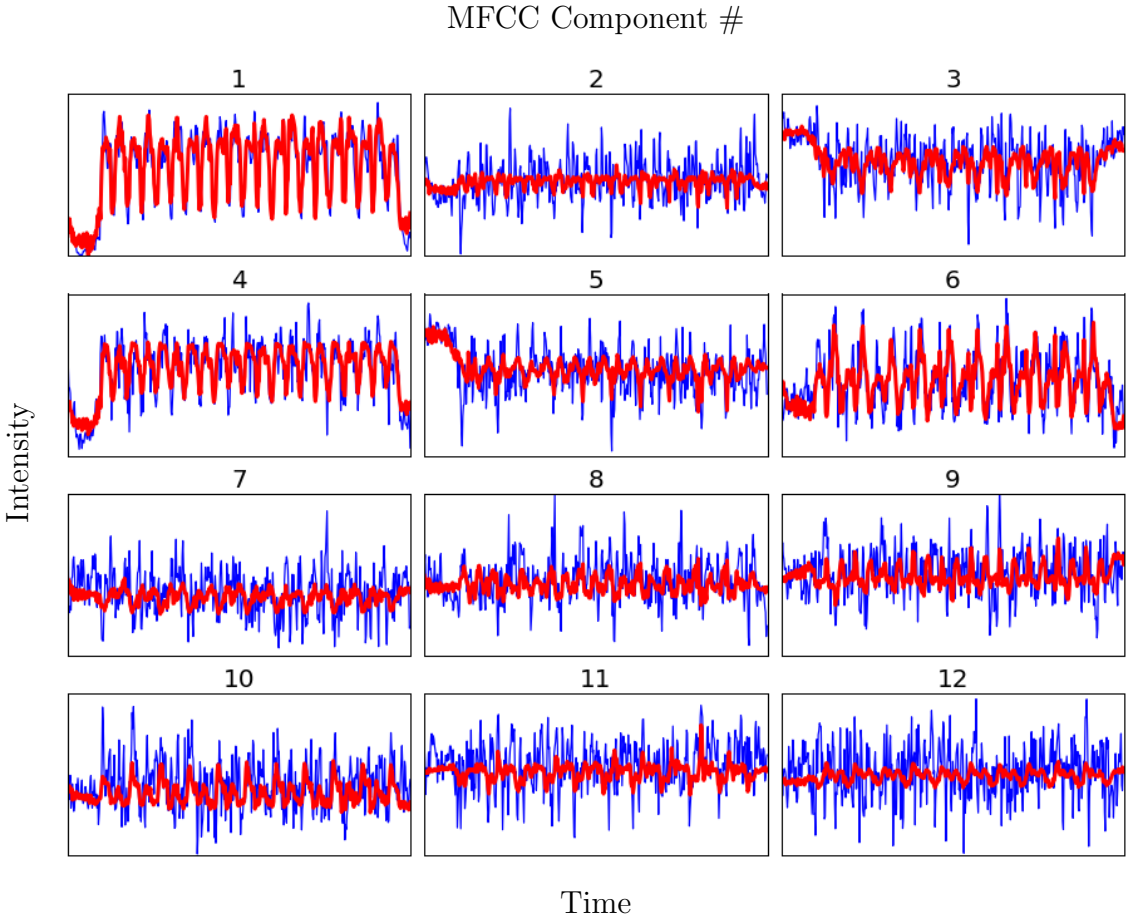

Figure 5: MFCC signals from a given trial (walk dataset) and the prediction of the neural network. MFCC components 1-12 are shown. Blue lines: actual MFCC signal. Red lines: Network prediction.

	Carpet	Desk	Tiles
Carpet	.454	.198	.318
Desk	.212	.500	.260
Tiles	.332	.298	.424

Table 2: Classification performance for predicting surface type in the walking experiment. Diagonal values show correct performance for each class. Values are averaged across 5-fold crossvalidation. Inside a crossvalidation fold, data (2500 samples for each class) was divided as %80 for training (2000 samples), and %20 for testing (500 samples).

Although prediction and subtraction of the walking signal was achieved (Figures 5

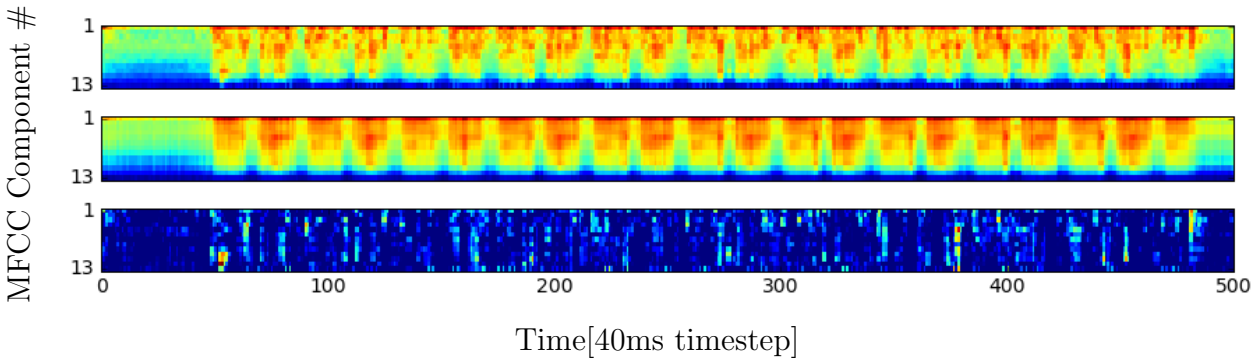

Figure 6: Demonstration of the signal subtraction process for one trial in the walking experiment. MFCC signals are converted into the log-filterbank format, which allows for subtraction. Upon subtraction, spectral flooring is applied to negative log values. Up: Heatmap of the original signal. Middle: Heatmap of the predicted signal. Down: Subtracted Signal.

and 6), our attempts at classification (Table 2) have been unsuccessful. Specifically, surface classification performance was at %42-50 in comparison to a chance level of %33. To improve classification performance, two possible solutions were implemented; namely, (i) abandoning the MFCC compression, and (ii) low-pass filtering the ego noise.

Using identically sized filterbanks for compression

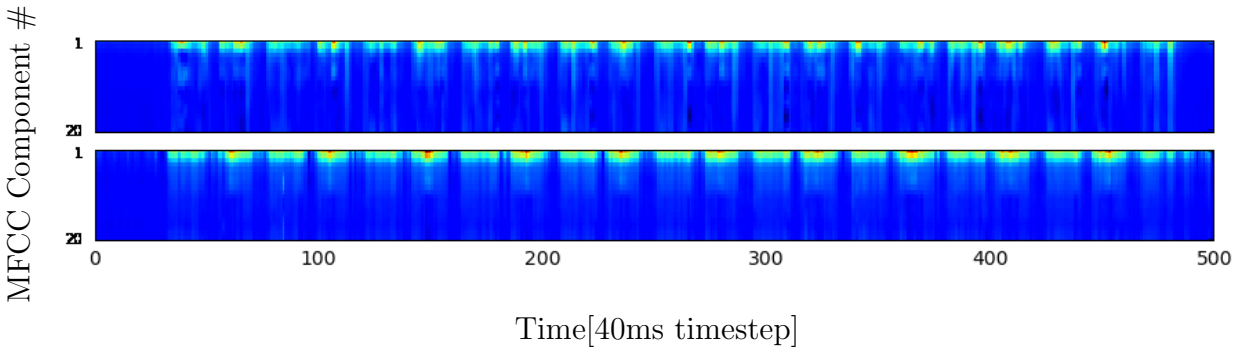

Figure 7: Heatmaps of the identical filterbank dataset and network predictions for this dataset. Since the log-filterbank format does not apply to this compression, it is not applied, and the signals are not subtracted. Up: Heatmap of the original signal. Down: Heatmap of the predicted signal.

The first approach was to use an alternative to the MFCC compression. In MFCC compression, lower frequency domains are assigned much narrower filterbanks (inside which all values are summed up together) than higher frequency components. Thus, MFCC compressions highly favor low-frequency information and sum up high-frequency domains in large bins. This can cause loss of information, as an important high-frequency source could be blended out with large amounts of noise in the summing process.

Thus, we initially used a compression with identically sized filterbanks instead of MFCC. Figure 7 shows the network predictions compared to the real data in the iden-

tical filterbank compression. Compared to the MFCC compression, it can be seen that irregular perturbations are more common. Classification of surfaces were at chance level with this approach.

Low-pass filtering the ego-noise

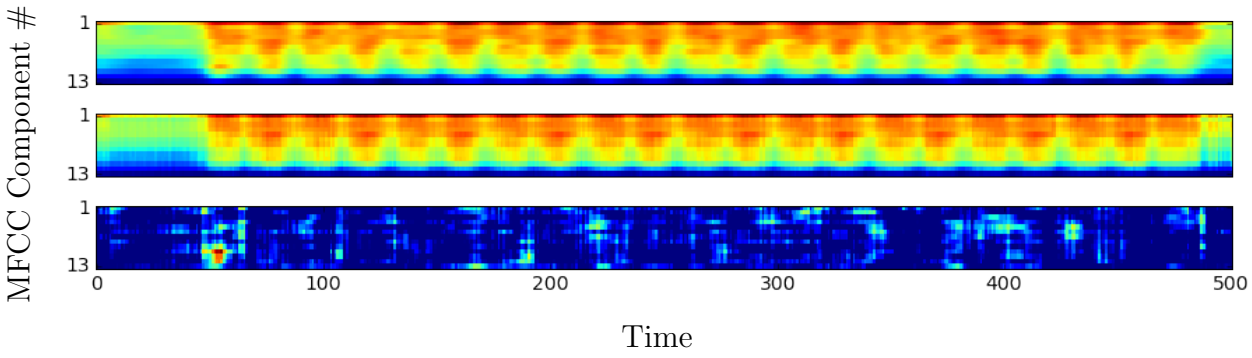

Figure 8: Demonstration of the signal subtraction process for one trial in the low-pass filtered walking dataset. MFCC signals are converted into the log-filterbank format, which allows for subtraction. Upon subtraction, spectral flooring is applied to negative log values. Up: Heatmap of the original signal. Down: Heatmap of the predicted signal.

	Carpet	Desk	Tiles
Carpet	.596	.164	.328
Desk	.118	.642	.188
Tiles	.290	.196	.494

Table 3: Classification performance for predicting surface type in the walking experiment. Diagonal values show correct performance for each class. Values are averaged across 5-fold crossvalidation. Inside a crossvalidation fold, data (2500 samples for each class) was divided as %80 for training (2000 samples), and %20 for testing (500 samples).

The second approach to ameliorate classification performance was to low-pass filter the signals in each MFCC channel to reduce perturbations that were unrelated to the walking motions. Low-pass filtering was achieved by convolving the MFCC channels with a Gaussian kernel ($SD \equiv 3$). Figure 8 and Table 3 show prediction and classification performance, respectively. Surface classification performance appears to have been improved by %7-15 depending on surface category.

Discussion

The results above demonstrate that forward models can be used to predict ego-noise reliably both on tasks involving one or multiple joints. This allows for the subtraction of the ego-noise to enhance signals from the environment.

The classification across different surfaces in the walking data, although above chance, produced markedly worse results than head motion data. One possible reason might have to do with the separability of the classes. The "desk" class was consistently classified

more accurately than "carpet" or "tile" classes. On the other hand, "carpet" and "tile" classes were mistaken as one another more often. These patterns are visible in both tables in the results section.

Another reason for the classification results could be the involvement of many more joints in the walking experiment in comparison to head motion experiment. In movement, the rotation of each joint creates rapid perturbations in each channel, which we identified as noise because they dont follow the cyclical pattern of the robots walking motion. Since we only used joint motions and positions, these signals were not predicted by our neural network algorithm. Two different solutions were implemented in order to address this noise.

Firstly, we hypothesized that implementing identically sized filterbanks could reveal channels that contain information about a given surface. Thus, taking a discrete Fourier transform of the raw sound signals, we produced a new dataset that has 20 channels with equally sized filterbanks. Analysis of this dataset, however, has returned chance-level classification rates.

The explanation for the failed results of this approach is likely in the classification procedure. The noise likely introduces large perturbations in the computed Euclidean distance of each network prediction. Each new introduced channel has an equal effect on the distance calculation. Therefore, if a lot of extra channels with low signal-to-noise ratio are introduced, which is what seems to be the case in this instance, the relative importance of the more informative channels will be reduced; leading to a performance decrease.

A second approach was then based on the idea of removing this noise. Although possibly also removing surface information, this would allow the training and classification algorithms to give more robust results. In line with this idea, the MFCC signals were low-pass filtered by convolution with a Gaussian kernel. Even though the signal subtraction results demonstrate that the network predictions are not far superior than the initial results, classification rates seem to be improved by a considerable amount. It could be argued that what the low-pass filtering does is "silencing" the channels that have non-relevant information to the walking motion. As above, a similar explanation could be attributed to our classification paradigm, in which the relative weight of the more informative channels are increased during distance calculation.

Future Directions

One of the initial goals of the lab rotation project was to implement this ego-noise attenuation system into an architecture using self-organizing maps. This approach is not only biologically more realistic (imitating motor mapping systems in living agents), but also advantageous in numerous ways, such as scalability, versatility of use and possibilities for on-line prediction making (Escobar-Juárez, Schillaci, Hermosillo-Valadez, & Lara-Guzmán, 2016). Unfortunately, the implementation and improvement of the above-mentioned aspects of the project proved to be too time consuming to implement this feature.

Although the high frequency perturbations are dismissed here as noise, they are simply a part of the ego-noise and are caused by the rotation of the joints. Therefore, this noise should theoretically also be predictable. Records of electric current given to the robots joints seem to follow similar rapid alterations in terms of signal intensity. Unfortunately, our work hasnt been able to successfully integrate the electric current information

to the MFCC predictions or surface classification. However, high-pass filtering the MFCC signals (since the low frequency variation seems to be explained well considering our findings) and then trying to predict the high perturbations using electric current recordings through forward models could be an interesting direction to take this project.

Finally, use of multiple data points in the analysis may be a promising way of improving classification results. In this project, we attempted to use joint information from multiple data points to predict a single MFCC vector, but this hasnt resulted in a remarkable improvement in classification.

Another approach that could be useful is to predict multiple MFCC vectors. The simplest example of this is to make one prediction out of multiple data points by picking the mode (the most frequently occurring prediction). For example, given a classification rate p $\in$ [0, 1], the misclassification rate would then be (1-p); and the misclassification rate for a prediction containing n data points should then be $(1\text{-p})^n$, which is exponentially smaller assuming a non-chance success rate. The issue with this approach is that it loses information, specifically, the distance values. A more elegant way could be to sum up the distance values of the predictions in question, but it is unclear whether this approach would be effective, especially if done without the low-pass filtering. Yet another option is to have the neural networks predict the MFCC vectors of multiple data points at once, but this will be computationally more intense.

To summarize, this project extensively used neural networks to resolve some simple problems of robot ego-noise. While ego-noise subtraction was carried out with relative success, surface classification tasks proved to be more challenging. Though we have managed to slightly improve our results, a plethora of applications remain to be tested for further amelioration of walking surface classification.

References

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357–366.

Escobar-Juárez, E., Schillaci, G., Hermosillo-Valadez, J., & Lara-Guzmán, B. (2016). A self-organized internal models architecture for coding sensory–motor schemes. *Frontiers in Robotics and AI*, *3*, 22.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive science*, *16*(3), 307–354.

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural networks*, *9*(8), 1265–1279.

Schillaci, G., Ritter, C. N., Hafner, V. V., & Lara, B. (2016). 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)* (pp. 390–397).