Prediction of Quadcopter State through Multi-Microphone Side-Channel Fusion

Hendrik Vincent Koops

Kashish Garg

Munsung (Bill) Kim

Jonathan Li

Anja Volk

Franz Franchetti

Technical Report UU-CS-2017-001
January 2017

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

ISSN: 0924-3275

Department of Information and Computing Sciences
Utrecht University
P.O. Box 80.089
3508 TB Utrecht
The Netherlands

Prediction of Quadcopter State through Multi-Microphone Side-Channel Fusion

Hendrik Vincent Koops[1], Kashish Garg[2], Munsung (Bill) Kim[2], Jonathan Li[2], Anja Volk[1], and Franz Franchetti[2]

[1]Department of Information and Computing Sciences, Utrecht University
[2]Department of Electrical and Computer Engineering, Carnegie Mellon University

Abstract

Improving trust in the state of Cyber-Physical Systems becomes increasingly important as more tasks become autonomous. We present a multi-microphone machine learning fusion approach to accurately predict complex states of a quadcopter drone in flight from the sound it makes using audio content analysis techniques. We show that using data fusion of multiple microphones, we can predict states with near-perfect results. Furthermore, we significantly improve the state predictions of single microphones, outperforming several other integration methods. These results show that side-channel information can be effectively used to improve the state assurance and security in Cyber-Physical Systems.

1 Introduction

Obtaining high-assurance state information for Cyber-Physical Systems (CPS) becomes increasingly important as more tasks become autonomous. A self-driving car that cannot accurately determine its own position, or a unmanned delivery drone flying in the wrong direction are examples of how incorrect state information can lead to catastrophic incidents and/or mis-delivered packages.

To determine the state of a vehicle, information from one or more primary sensor measurements such as speedometers, accelerometers or GPS is most often used. However, it has been shown that most, if not all sensors are susceptible to spoofing attacks. This problem is accelerated with the trend of increased sensors connectivity with (wireless) networked systems and the Internet [1]. In spoofing, real sensor data is maliciously replaced by falsified data in an undetectable way. For GPS for example, this can result in a false estimation of position. Over the years, multiple solutions to improve the security of the sensors have been proposed. Most of these rely on cryptographic solutions or data analysis of the sensor signal to find anomalies that are indicative of falsification [2]. In addition to the obvious importance of primary sensor security improvement, state estimation can also be improved by using side-channel information from sensors that use data from a different domain.

Using cues from different domains to establish a (high) trust in a state is similar to how humans assess their environment. For instance, when we drive a car at a constant speed, and the sound of the engine suddenly changes while the speed does not, we instantly assume that the difference is indicative of something being wrong with the vehicle. It has been shown that multicomponent (or multi-modal) signals can improve error detection by producing multiple components together, thus reducing the reaction time, increasing the probability of detection and lowering the intensity at which detection occurs [14]. We can use the intuition of multi-modal interaction in CPS to improve the security and assurance of primary sensors measurements, by using side-channel information to make a state prediction. For example, recent research has shown that audio data from a microphone can be used to estimate various states (i.e. speed and gear position) from a moving vehicle with high precision [10]. Thus, reliable side-channel information can be used to improve CPS security by matching and checking the validity of the data from the primary sensor measurements.

Research into using audio information for the analysis of physical systems is not new. Nevertheless, nearly all of this research is aimed at detecting low-level system states, such as malfunction or fault detection of engines, gears or bearings [5,13]. We propose a more complex analysis and diagnosis of the sound a CPS produces. For this, we take inspiration from a research area related to digital signal processing called audio content analysis. This type of audio research in Music Information Retrieval (MIR) aims to analyze musical audio signals to analyze, detect or estimate concepts with various degrees of abstraction in the audio signal that stem from musicological research. In addition to research into lower level concepts such as notes, (etc), MIR is actively researching more complex concepts such as chords, genre and emotion. Similarly, instead of just detecting whether a drone behaves faulty compared to a baseline measurement, we aim to solve the task of predicting more complex states a quadcopter can be in, such as hovering, descending and ascending from the sound it makes. To this aim, we propose to use the audio from multiple microphones.

Recent research in data integration has shown that information from multiple heterogeneous sources can be integrated to create improved, and more reliable data [4]. These *data fusion* techniques have for example been successfully applied in MIR in a musical context of integrating crowd-sourced chord sequences [9]. It was shown that integrated data outperforms individual source data, and that it can be used to accurately estimate the relative quality of sources. Therefore, instead of using just one microphone side-channel, we propose to integrate the state estimations of four microphones, each recording the sound of a single rotor of a quadcopter. We show that using multiple microphones, we can predict with near-perfect precision whether a quadcopter is either descending, hovering or ascending.

Contribution The contribution of this paper is threefold. Research in audio content analysis in MIR often takes inspiration from other fields, but the reverse is rare. First, with techniques inspired from MIR, we show that we can predict complex state information from the sound a quadcopter makes. Secondly, we show how predictions from multiple microphones can be integrated to obtain an improved prediction using data fusion. Thirdly, we show how using data fusion, we can improve CPS security by accurately estimating the relative quality of a source.

Synopsis The remainder of this paper is structured as follows. Section 2 introduces our method of improving quadcopter state security by predicting and integrating machine learning outputs from multiple microphone sources. Section 3 provides results of these integration methods, and Section 4 closes with conclusions.

2 Quadcopter state prediction and integration from audio

This section details the method used to integrate multiple predictions of the state of a quadcopter from the sound its rotors make during flight. To achieve this, we fly a 3DR IRIS+ quadcopter a predefined flight plan in autopilot mode, while four microphones attached to each of the four arms of the quadcopter record the sound of the rotors. During flight, an on board computer records ground truth information about the state of the quadcopter, as detailed in Section 2.1.1. From the audio of each of the microphones, we extract features that are used in a machine learning classification task, as detailed in Section 2.1.2. We classify the audio features of each of the microphones individually, as detailed in Section 2.2. To improve the classification results of the individual microphones, we integrate their predictions, as detailed in Section 2.2.1.

2.1 Data collection

The flight of a quadcopter is easily influenced by weather conditions such as wind, and the pilot (controller) by means of overcompensation. To control for pilot influence during flight, we set up a controlled environment where the quadcopter is flying a preprogrammed path in autopilot mode. The flight plan consists of seven steps:

1. Take-off and ascend to 5 meters
2. Hover for 10 seconds
3. Ascend to 10 meters
4. Hover for 10 seconds
5. Descend to 5 meters
6. Hover for 10 seconds
7. Descend and land

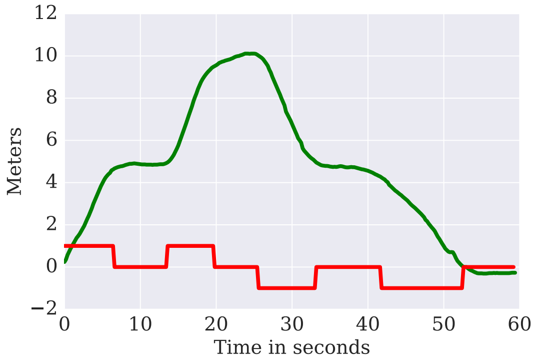

Figure 1: Example of recorded altitude of autopilot flight plan in green and derived ascending (1), hovering (0) and descending (-1) data in red.

A visualization of this flight path can be found in green in Figure 1. Flights were performed in an open field in dry weather conditions. During the autopilot flight, we recorded both ground truth state information and the sound of each of the rotors.

2.1.1 Telemetry

During flight, we used the on-board telemetry system to record quadcopter flight data at a fixed sampling frequency. In this research, we focus on predicting three states of the quadcopter during flight: ascending, hovering and descending (AHD). The quadcopter itself does not record this data, but it does record data from which we can derive these states, i.e. the absolute altitude measured by the on-board GPS receiver. To calculate AHD data, we compute a gradient from the altitude information, from which we calculate a step function that describes if the gradient is increasing, stable or decreasing, which we interpret as AHD.

The gradient is calculated through a first-order discrete difference along the list of altitudes. Suppose we measure the altitude at a certain sampling frequency to be $[0, 5, 5, 5, 0]$, that is: starting at 0 meter followed by 3 frames at 5 meter and finally back at 0 meter again. Computing the gradient results in differences $[5, 0, 0, -5]$, from which we only keep the sign of the numbers and the zeros. The result of this example is $[+1, 0, 0, -1]$, which we interpret as $+1$, 0 and -1 as an ascending state, hovering state and descending state, respectively. An example of derived AHD state information from GPS altitude data can be found in red in Figure 1.

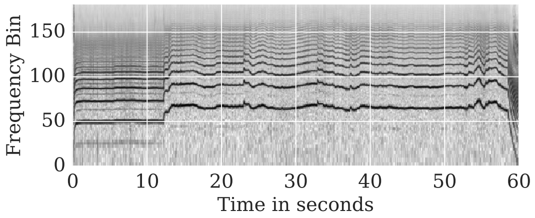

Figure 2: Example of a spectrogram of rotor audio, while a quadcopter performs the sequence described in Section 2.1 and Figure 1

2.1.2 Rotor audio feature extraction

To record the sound of the quadcopter's rotors during flight, we equipped the quadcopter with four microphones, one underneath each of the arms, close to the rotors. We record the sound at 44.1 kHz, 16-bit. As Figure 2 shows, the rotor sound is rich in content at higher frequencies. Therefore, the audio is passed through a nonuniform filter bank of 24 bands per octave to create a spectrogram that increases in detail with frequency. From this filtered signal, we create a logarithmically filtered short-time Fourier transform spectrogram at ten frames per second with a frame size of 8192 samples, with a minimum and maximum frequency of 30Hz and 60kHz, respectively. From a visual inspection of the spectrogram, this frequency range was found to have the most important information. From preliminary experiments, it was found that frequency analysis beyond these bounds did not significantly improve results. Nonuniform filtering and short-time Fourier transform results in a spectrogram representation of the signal in 181 bins per audio frame.

2.1.3 Context window

Detection of more complex events in audio content analysis often improves with the use of a context window around an audio frame. Research in automatic chord estimation [11] and speech recognition [6] provide examples where context windows have proven to be successful in improving classification.

Therefore, we concatenate consecutive frames of the spectrogram within a context window to form the input for a classifier. More specifically, for a context window W_i of size of n, we concatenate the frames $f_{i-n/2}$ to $f_{i+n/2}$ to classify frame f_i from the spectrogram. We experiment with different window sizes to find the optimal amount of context in terms of classification accuracy. These concatenated spectrogram frames are used as input for a classifier.

2.2 Classification

Although recent advances in deep learning have shown great results in learning high level abstractions from audio, we choose a fast, lightweight solution that in theory can run from an on-board quadcopter computer in real-time. From a preliminary experiment, it was found that a Random Forest Classifier (RF) produced the best results from a selection of learning algorithms. RF [7, 8] is an ensemble classifier that uses unpruned classification trees created from bootstrap samples of the training data and random feature selection in tree induction. Prediction is made by aggregating (majority vote or averaging) the predictions of the ensemble. For a detailed description of RF we refer to [7, 8].

The context window frames of each of the microphones are classified using RF, resulting in four heterogeneous classification streams. An example of this can be found in Table 1, where the classification results of four consecutive context windows from the four microphones $M_0 \ldots M_3$ can be found. To take advantage of the shared information between the microphones, we propose to integrate the classifications using data fusion.

2.2.1 Integration

To find the best state prediction from four individual microphones, we propose to use data fusion. We compare the results of data fusion with three baseline methods: majority voting, random picking and the average microphone accuracy.

Random picking (RND) selects a state from a randomly picked microphone for every context window. For the example in Table 1, RND essentially picks one state from 4^4 possible state combinations by picking a state from a randomly chosen microphone per context window.

	W_i	W_{i+1}	W_{i+2}	W_{i+3}
M$_0$	Ascend	Descend	Descend	Hover
M$_1$	Ascend	Hover	Hover	Hover
M$_2$	Ascend	Hover	Descend	Descend
M$_3$	Ascend	Hover	Descend	Descend
MV	Ascend	Hover	Descend	?
DF	Ascend	Hover	Descend	Descend

Table 1: Example of four microphones M($_{0...3}$) providing different state classes (Ascend, Hover, and Descend) for a sequence of context windows. DF shows an example output of data fusion on these sources. DF is identical to majority vote (MV) on the first three states. For the last states, fusion chooses Descend by taking into account source accuracy, while majority vote would randomly pick either Hover or Descend.

Majority voting (MV) selects the most frequent state shared between the microphones for every context window. In case multiple states are most frequent, we randomly pick from the most frequent states. For the example in Table 1, the chosen states would be either Ascend, Hover, Descend, Hover or Ascend, Hover, Descend, Descend.

Data Fusion (DF) can be viewed as an extension of majority voting in the sense that in addition to finding the most common symbol per audio frame, it also uses the agreement between microphones. Microphones with higher agreement with other microphones are considered to be more trustworthy. We propose to a method adapted from ACCUCOPY model introduced by Dong et al. in [3, 4] to integrate conflicting databases. Instead of databases, we propose to integrate state predictions. This adapted model was previously successfully applied in a musical context, where it showed to outperform majority voting and random picking in an automatic chord extraction task [9].

Calculating the data fusion integration happens in two steps: after computing the state class likelihoods for each context window per microphone, a *source accuracy* is computed for each microphone by taking the mean of all its state likelihoods. The likelihoods of each state are then weighted by their source accuracy. The intuition here is that microphones with higher agreement with other microphones are more trustworthy. The process of computing state likelihoods and source accuracy is repeated until the likelihoods of the states converge. For each audio frame, the value with the highest likelihood is chosen. For a detailed description of data fusion, the reader is forwarded to [3, 4, 9].

Average accuracy (AVG). To assess the improvement over the average microphone in terms of classification accuracy, we also compare the results of DF, RND and MV with the average classification accuracy of the microphones. This will show how much on average the integration methods will improve the classification results of the individual microphones.

Another intuitive way of integrating the classifications of RF would be to average or multiply the class probabilities of every microphone per context window. Nevertheless, from preliminary experiments it was found that this scales quite poorly. With an increased amount of classes, averaging or multiplying the class probabilities resulted in accuracies nearing random classification accuracy. This is because with an increased number of classes, the joint-probability has a higher probability of going towards a uniform distribution. On the other hand, research in several domains shows that data fusion scales very nicely [4, 9, 12].

2.3 Cross-validation

To test how well our method generalizes, we cross-validate our method on multiple iterations of the same flight plan. The quadcopter executes the flight plan as mentioned in Section 2.1.1 15 times. For every flight, and every context window size we perform 20-fold cross validation on a randomly selected 70-30% train/test set split of our dataset. For each of the integration methods, we take the average of the cross validation over all flights as the final classification accuracy.

3 Results

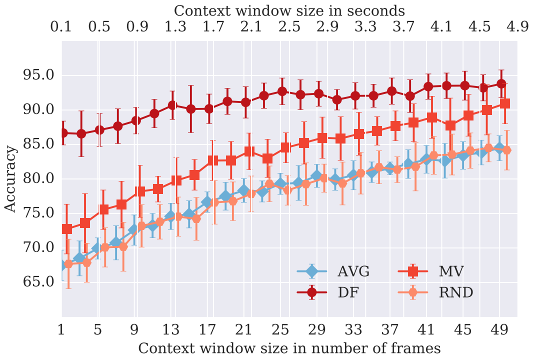

Figure 3: 20-fold cross-validation classification results for DF, MV, RND of integrating the classifications of four microphones for several context window sizes.

Classification integration results for several context window sizes for DF, MV and RND can be found in Figure 3. The figure shows that RND does not improve the average microphone classification, performing equally with the average microphone at every context window size. MV improves the average microphone accuracy with 5.5 percentage points on average for every context size. DF improves the average microphone accuracy the most, outperforming all other integration methods by 10 to 20 percentage points. For every context window size DF performs significantly better than MV with $p << 0.01$ using a Wilcoxon signed-rank test for the null hypothesis that two related paired samples come from the same distribution.

Furthermore, Figure 3 shows that classification results for all methods improve with context window size. DF seems to be more robust to this effect. Increasing the context window size increases the reaction time of the system: if more frames are needed to make a good state estimation, more time is needed. Therefore, the smaller the frame size the better. We find that DF integration stabilizes at around context windows sizes of 13 frames. For the other integration methods, we find that accuracy increases almost linearly with context window size. This shows that DF is capable of finding useful shared knowledge between the microphones to make a good integration.

4 Conclusions

We have shown that through audio content analysis of quadcopter drone rotor audio, we can accurately predict states a quadcopter is in. More specifically, we have shown that through integration of classifications from multiple microphones, we can significantly improve the prediction of complex states that describe whether a quadcopter is ascending, hovering or descending.

Our research expands earlier research on state prediction of Cyber-Physical Systems from the sound they make [10]. Furthermore, it shows the benefit of multi-modal state estimation by using side-channel information from a different domain than primary sensors use. Our results show that side-channel information can be used to obtain complex state descriptions with high accuracy. Therefore, we predict that future research will show the benefit of improving the security of CPS by estimating complex states from a multitude of side-channel sources. We believe that combining the estimations from a large number of side-channel sources working on different domains can greatly improve the security of many CPS.

Acknowledgment

H.V. Koops and A. Volk are supported by the Netherlands Organization for Scientific Research, through the NWO-VIDI-grant 276-35-001 to A. Volk. This material is based on research sponsored by DARPA under agreement number FA8750-12-2-0291. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

References

[1] Alvaro A Cardenas, Saurabh Amin, and Shankar Sastry. Secure control: Towards survivable cyber-physical systems. *System*, 1(a2):a3, 2008.

[2] Yingying Chen, Wenyuan Xu, Wade Trappe, and YanYong Zhang. *Securing Emerging Wireless Systems: Lower-layer Approaches*. Springer Science & Business Media, 2008.

[3] X.L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *Proc. of the VLDB Endowment*, 2(1):550–561, 2009.

[4] X.L. Dong and D. Srivastava. Big data integration. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 1245–1248. IEEE, 2013.

[5] Patricia Henriquez, Jesus B Alonso, Miguel A Ferrer, and Carlos M Travieso. Review of automatic fault diagnosis systems using audio and vibration signals. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(5):642–652, 2014.

[6] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[7] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[8] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

[9] Hendrik Vincent Koops, W Bas de Haas, Dimitrios Bountouridis, and Anja Volk. Integration and quality assessment of heterogeneous chord sequences using data fusion. In *International Society for Music Information Retrieval Conference*, pages 178–184, 2016.

[10] Hendrik Vincent Koops and Franz Franchetti. An ensemble technique for estimating vehicle speed and gear position from acoustic data. In *Digital Signal Processing (DSP), 2015 IEEE International Conference on*, pages 422–426. IEEE, 2015.

[11] Filip Korzeniowski and Gerhard Widmer. Feature learning for chord recognition: The deep chroma extractor. In *Proc. of the 17th Int. Society for Music Information Retrieval Conf.(ISMIR)*, 2016.

[12] Xian Li, Xin Luna Dong, Kenneth B Lyons, Weiyi Meng, and Divesh Srivastava. Scaling up copy detection. In *2015 IEEE 31st International Conference on Data Engineering*, pages 89–100. IEEE, 2015.

[13] Marwan Madain, Ahed Al-Mosaiden, and Mahmood Al-khassaweneh. Fault diagnosis

in vehicle engines using sound recognition techniques. In *Electro/Information Technology (EIT), 2010 IEEE International Conference on*, pages 1–4. IEEE, 2010.

[14] Candy Rowe. Receiver psychology and the evolution of multicomponent signals. *Animal Behaviour*, 58(5):921–931, 1999.