2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2017)
November 16 - 18, 2017, Daegu, Korea

Multirotor UAV State Prediction through Multi-Microphone Side-Channel Fusion

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

***Abstract*— Improving trust in the state of Cyber-Physical Systems becomes increasingly important as more Cyber-Physical Systems tasks become autonomous. Research into the sound of Cyber-Physical Systems has shown that audio side-channel information from a single microphone can be used to accurately model traditional primary state sensor measurements such as speed and gear position. Furthermore, data integration research has shown that information from multiple heterogeneous sources can be integrated to create improved and more reliable data. In this paper, we present a multi-microphone machine learning data fusion approach to accurately predict ascending/hovering/descending states of a multi-rotor UAV in flight. We show that data fusion of multiple audio classifiers predicts these states with accuracies over 94%. Furthermore, we significantly improve the state predictions of single microphones, and outperform several other integration methods. These results add to a growing body of work showing that microphone side-channel approaches can be used in Cyber-Physical Systems to accurately model and improve the assurance of primary sensors measurements.**

I. Introduction

Obtaining high-assurance state information (such as speed, location, direction, etc.) from Cyber-Physical Systems (CPS) becomes increasingly important, especially as more of their tasks become autonomous. A self-driving vehicle 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 CPS state, such as speed, acceleration or location of CPS, we commonly rely on information from one or more primary sensor measurements such as speedometers, accelerometers or GPS. However, it has been shown that most, if not all sensors are susceptible to attacks. This problem is accelerated by the trend of increased sensors connectivity with (wireless) networked systems and the Internet [1], [2], [3], [4]. For GPS for example, this can result in a false estimation of position. Multiple solutions to improve trust in sensor information have been proposed. Most of these rely on improving sensor security through cryptographic solutions or data analysis of the sensor signal to find anomalies that are indicative of falsification [5].

[1]Hendrik Vincent Koops and Anja Volk are with the Department of Information and Computing Sciences, Utrecht University, the Netherlands {h.v.koops}{a.volk}@uu.nl

[2] Kashish Garg, Munsung (Bill) Kim, Jonathan Li and Franz Franchetti are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, USA {k.garg}{m.kim}{j.li}{franzf}@ece.cmu.com

Fig. 1. Quadcopter equipped with microphones (positioned in the red circles) close to each rotor. After classifying the sound of each rotor separately, we fuse the classifier outputs to obtain improved ascend/hover/descend state predictions.

A different approach to improving trust in CPS state estimation is to use side-channel information from sensors that use data from a different domain. Such state estimations from side-channel information can be used to verify or enrich primary sensor state estimations. For example, in a previous study it was shown that microphone audio can be used to accurately estimate states of a moving vehicle [6]. These side-channel estimations can be used in conjunction with primary sensors to improve the trust of state estimation.

In this paper, we introduce an approach to multi-microphone state prediction of a quadcopter drone in flight. More specifically, we investigate the use of multiple microphones as side-channel sensors for state prediction. We show that using multiple microphones, we can predict with near-perfect accuracy whether a quadcopter is either descending, hovering or ascending. Furthermore, we show that using a data fusion technique, we can accurately assess the relative quality of microphone data, by investigating their deviation from the consensus between the microphones.

Contribution. The contributions of this paper are as follows. First, we show that we can predict ascending, descending and hovering states from the sound a quadcopter makes. Secondly, we show how predictions from multiple microphones can be integrated to obtain an improved state prediction using data fusion. Thirdly, we show we can accurately estimate the relative quality of microphone side-channel data using data fusion.

Synopsis. The remainder of this paper is structured as follows. Section III introduces our method of improving quadcopter state assurance by predicting and integrating machine learning outputs from multiple microphone sources. Section IV details the way we evaluate our system. Section V provides results of these integration methods, and Section VI closes with conclusions.

978-1-5090-6064-1/17/$31.00 ©2017 IEEE

II. Background

To establish high trust in a state, humans assess their environment using information from different domains. For example, assume we drive a vehicle at a constant speed, and the speedometer suddenly indicates zero while the sound of the engine stays the same. We will immediately assume that the speedometer is faulty and trust our other senses that the vehicle is still driving at the same speed. We can use this intuition in CPS by using side-channel information to make state predictions. For example, research has shown that audio data from a microphone can be used to accurately estimate various states (i.e. speed and gear position) from a moving vehicle with [6].

Audio state estimation of CPS. Research into using sound 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 [7], [8]. We propose to estimate more complex states from the sound a CPS produces. For this, we take inspiration from a research area related to digital signal processing called audio content analysis. Instead of just detecting whether a CPS behaves faulty compared to a baseline measurement from the sound it makes, we propose to investigate more complex states. Some examples are detecting the state a quadcopter can be in, for example hovering, descending and ascending.

Data fusion. Recent research in data integration has shown that information from multiple heterogeneous sources can be integrated to create improved, and more reliable data [9]. These *data fusion* techniques have for example been successfully applied in a musical context of integrating crowd-sourced chord sequences [10]. It was shown that integrated data outperforms individual source data, and that it can be used to accurately estimate the relative quality of data sources. In this research, we will combine CPS state estimation from audio with data fusion, to improve the accuracy of state prediction, and therefore obtain a higher trust in state estimations from sound.

III. 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. 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 state information, as detailed in Section III-A.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 III-A.2. We classify the audio features of each of the microphones individually, as detailed in Section III-B. To improve the classification results of the individual microphones, we integrate their predictions, as detailed in Section III-C.

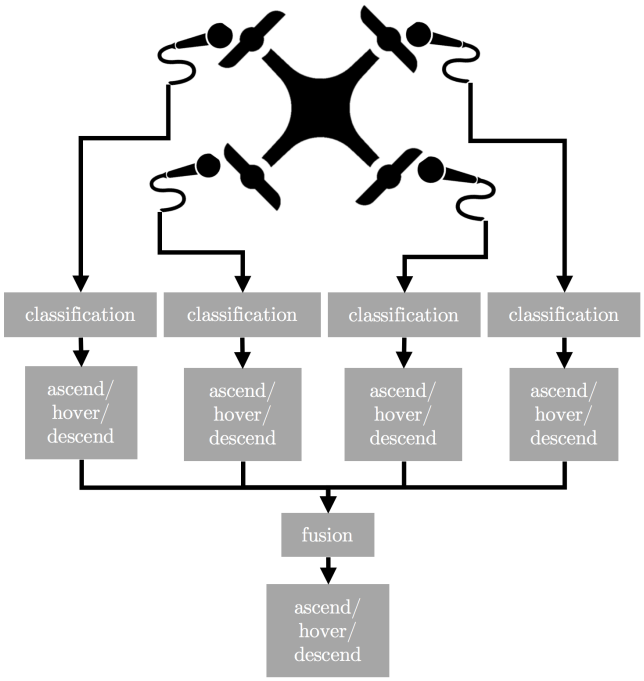

Fig. 2. Schematic pipeline of the proposed system. The sound of each rotor is classified using a Random Forest classifier. The outputs of all classifiers are then integrated using Data Fusion to create an improved state prediction.

A. 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

A visualization of this flight path can be found in green in Fig. 3. Flights were performed in an open field in dry weather conditions. During the autopilot flight, we recorded both ground truth state information through telemetry (Section III-A.1) and the sound of each of the rotors (Section III-A.2).

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.

Obtaining AHD. To calculate AHD, we compute a gradient from the altitude data, from which we calculate a step function that describes if the gradient is increasing, stable

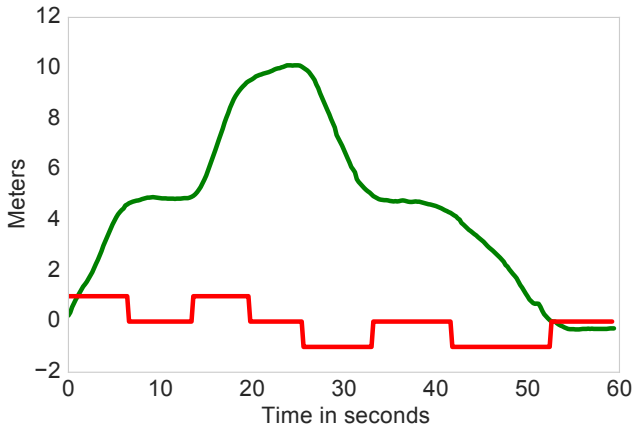

Fig. 3. Example of recorded altitude of autopilot flight plan in green and derived ascending (1), hovering (0) and descending (-1) data in red.

or decreasing. We interpret this step function as AHD. The gradient is calculated through a first-order discrete difference of the sampled 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 samples 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 Fig. 3.

2) Rotor audio feature extraction: To record the sound of each quadcopter rotor during flight, we equip the quadcopter with four microphones, one above each of the arms, close to the rotors. We record the sound at 44.1 kHz, 16-bit. As the spectrogram of an example recording in Fig. 4 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 increases frequency content detail at higher frequencies.

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 18kHz, 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.

3) Context window: Research in a large number of audio content analysis experiments has shown that better prediction accuracies can be achieved by aggregating information over several frames instead of using a single frame. Research in speech recognition [11] and automatic chord estimation [12] provide examples where context windows have proven to be successful in improving classification.

Therefore, we concatenate consecutive frames (context window) of the spectrogram to form the input to a classifier. More specifically, to classify frame f_i from the spectrogram, we concatenate the frames $f_{i-n/2}$ to $f_{i+n/2}$ to create a context window W_i of size n, where $n \in 2\mathbb{N}_{>0}$. These concatenated spectrogram frames are used as input for a classifier. We experiment with different window sizes to find the optimal amount of context in terms of classification accuracy.

B. Classification

Although recent advances in deep learning have shown great results in machine learning using deep architectures, 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 the commonly used Random Forest Classifier (RF) produced the best results from a selection of learning algorithms. RF [13], [14] 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, thereby creating a strong classifier from multiple weaker ones. It is beyond the scope of this paper to fully describe RF. For a complete description we refer to [13], [14].

The context window frames of each of the four microphones are classified using RF, resulting in four heterogeneous classification streams. An example of this can be found in Table I, where the classification results of four consecutive context windows from the four microphones M_0, M_1, M_2 and M_3 can be found. We hypothesize that the shared information between the microphones can be used to improve the classification accuracy over using a single microphone. To integrate the shared information between the microphones, we propose to use data fusion and compare its results with other integration methods.

C. Integration

To find the best state predictions among the classification results of four individual microphones, we explore several integration methods. We compare the baseline methods *random picking* and *majority voting* with *data fusion* integration, and compare them with the average microphone accuracy.

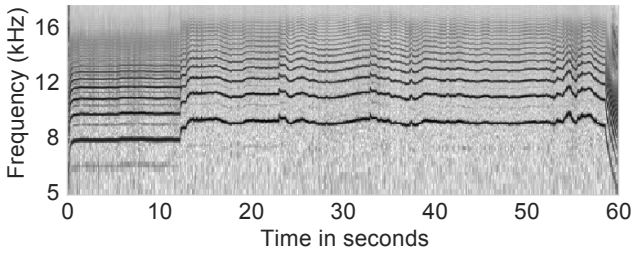

Fig. 4. Example of a spectrogram of rotor audio, while a quadcopter performs the sequence described in Section III-A and Fig. 3

TABLE I

STATES (ASCEND, HOOVER, AND DESCEND) PREDICTED FROM A SEQUENCE ($W_i \ldots W_{i+3}$) OF AUDIO CONTEXT WINDOWS FROM FOUR MICROPHONES M$_{(0\ldots3)}$.

	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

*1) Random Picking (*RND*):* selects a state from a randomly picked microphone for every context window. For the example in Table I, RND essentially picks one state from 3^4 possible state combinations by picking a state from a randomly chosen microphone per context window.

*2) 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 I, MV would result in either `Ascend, Hover, Descend, Hover` or `Ascend, Hover, Descend, Descend`.

*3) Data Fusion (*DF*):* can be viewed as an extension of majority voting in the sense that in addition to finding the most common state per audio context window, it also uses the agreement between microphones to integrate data. 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 [9], [15] to integrate conflicting databases. This model was previously successfully applied in a musical context, where it showed to outperform baseline methods in an automatic chord extraction task [10]. In this study, we propose to integrate RF state predictions from four microphones. In the following sections, we refer to the RF output of a single microphone as a *source*, which provides a sequence of state predictions. Calculating DF involves the computation of *source accuracy*, *vote counts*, and *state probabilities*.

Source accuracy is calculated by taking the arithmetic mean of the probabilities of all states the source provides. As an example, suppose we estimate the probabilities of the states in Table I based on their frequency count (c.q. likelihood). That is, `Ascend` for the first column is 1, `Descend` for the second column is $1/4$, etc. Taking the average of the state probabilities of the first source in our example of Table I we can calculate the source accuracy $A(M_0)$ of M_0 as follows:

$$A(M_0) = \frac{1 + 1/4 + 3/4 + 1/2}{4} = 0.625 \tag{1}$$

In the same way, we can calculate the source accuracies for the other three sources which are 0.625, 0.75 and 0.75 for M_1, M_2 and M_3 respectively.

Assuming that the sources are independent, then the probability that a source provides a correct state is its source accuracy. Conversely, the probability that a source provides an incorrect state is the fraction of the inverse of the source accuracy over all possible incorrect values n: $\frac{(1-A(M))}{n}$. In our case $n = 2$, since we have three possible states. The states of sources with higher source accuracies are more likely to be selected through the use of vote counts.

Vote counts are used as weights for the probabilities of the states they provide. With n and $A(M_i)$ we can derive a vote count $VS(M_i)$ of a source M_i. The vote count of a source is computed as follows:

$$VS(M_i) = \ln \frac{nA(M_i)}{1 - A(M_i)} \tag{2}$$

Applied to our example, this results in vote counts of 2.62 for M_0 and M_1, and 2.80 for M_2 and M_3. The higher vote count for M_2 and M_3 means that its values are more likely to be correct than those of M_0 and M_1.

State probabilities: After having defined the accuracy of a source, we can now determine which states provided by all the sources are most likely correct, by taking into account source accuracy. In the computation of state probabilities we take into account a) the number of sources that provide those states and b) the accuracy of their sources. With these values we calculate the vote count $VC(\mathscr{L})$ of a state $\mathscr{L}$, which is computed as the sum of the vote counts of its providers:

$$VC(\mathscr{L}) = \sum_{\sigma \in S^{\mathscr{L}}} VS(\sigma) \tag{3}$$

where $S^{\mathscr{L}}$ is the set of all sources that provide the state $\mathscr{L}$. For example, for the vote count of `Hover` in the last column of the example in Table I, we take the sum of the vote counts of M_0 and M_1. For the vote count of `Descend` we take the sum of the vote counts of M_2 and M_3. To calculate state probabilities from state vote counts, we take the fraction of the state vote count and the state vote counts of all possible states D:

$$P(\mathscr{L}) = \frac{\exp(VC(\mathscr{L}))}{\Sigma_{l \in D} \exp(VC(l))} \tag{4}$$

Applied to our example from Table I, we see that solving this equation for `Hover` results in a probability of $P(\texttt{Hover}) \approx 0.39$, and for `Descend` results in a probability of $P(\texttt{Descend}) \approx 0.56$. Instead of having to choose randomly as would be necessary in a majority vote, we can now see that `Descend` is more probable to be the correct state, because it is provided by sources that are overall more trustworthy.

Iterative computation. State likelihoods and source accuracy are defined in terms of each other, which poses a problem for calculating these values. As a solution, we initialize the state likelihoods with equal probabilities and iteratively compute state likelihoods and source accuracy until the state probabilities converge or oscillation of values is detected. The resulting state is composed of the states with the highest likelihoods.

For detailed Bayesian analyses of the techniques mentioned above we refer to [15], [16]. With regard to the scalability of data fusion, it has been shown that DF with source dependency runs in polynomial time [15]. Furthermore, [17] proposes a scalability method for very large data sets, reducing the time for source dependency calculation by two to three orders of magnitude.

4) Average Microphone 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. Computing AVG simply produces the non-weighted mean of the accuracies (i.e. the proportion of true correct classifications compared to the ground truth) of all four microphones. Comparing the integration methods with AVG will show how much on average the integration methods will improve the classification results of the average microphone.

IV. Evaluation

We evaluate our system of audio feature extraction, classification and data integration integration accuracy using cross-validation (Section IV-A). Furthermore, we investigate the Data Fusion source accuracy measure (Section IV-B).

A. Integration Accuracy

To evaluate the integration methods, we perform cross-validation on 15 different iterations of the flight plan mentioned in Section III-A.1. For each of the 4 microphones, for all 15 flights, we perform 20-fold frame-wise cross validation on randomly selected 70/30% train/test set splits of the shuffled data. For each fold, a RF classifier is trained on the training set of the folds and tested on the testing set of the folds. The output of RF on the test set of each the microphones is integrated using each integration method (DF, RND, and MV). The average accuracy of the 20 folds is reported as the classification accuracy of each integration method. We repeat this process for each context window size.

To evaluate the output quality of each integration method, we compare their accuracies (i.e. the fraction of correct classifications with regard to the ground truth). We also compare these scores with the average microphone accuracy (AVG) to see how much integration improves classification from an average microphone. These results show whether we can improve the trust in quadcopter state estimation by data fusing audio side-channel data from multiple microphones, compared to using a single microphone.

B. Data Fusion Source Accuracy

Research in other domains has shown that DF Source Accuracy can be used as a way to rank sources by their quality without having ground truth knowledge [9], [10]. To evaluate it in the context of CPS in this paper, we compare for each microphone its DF Source Accuracy with its ground truth accuracy. This reveals if DF Source Accuracy is useful for estimating microphone data quality from their agreement with the other microphones.

V. Results

A. Integration Accuracy.

Classification results for several context window sizes for the different integration methods DF, MV and RND can be found in Fig. 5. The figure shows that DF produces the best results of all integration methods, up to 94.2%,

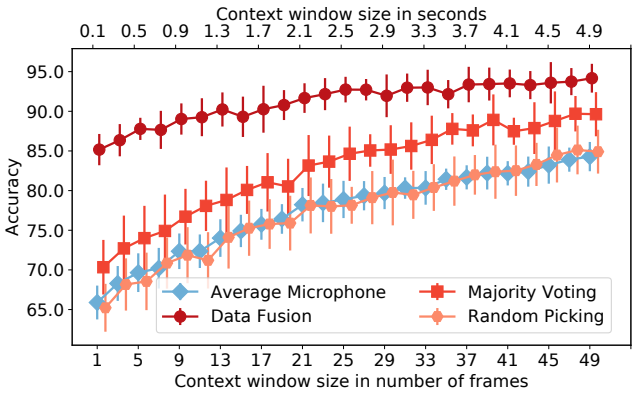

Fig. 5. Average 20-fold cross-validation classification results for Data Fusion (DF), Majority Voting (MV), and Random Picking (RND) of integrating the classifications of four microphones for several audio context window sizes. Average accuracy (AVG) shows the average microphone accuracy.

outperforming all other integration methods by around 10 to 20 percentage points. For every context window size DF performs significantly better than MV with $p \ll 0.01$ using a Wilcoxon signed-rank test for the null hypothesis that two related paired samples come from the same distribution [18]. MV improves the average microphone accuracy with 5.5 percentage points on average for every context size. RND does not improve the average microphone classification, performing equally with the average microphone at every context window size.

Effect of context window size. Fig. 5 shows that classification results for all methods improve with context window size, but DF seems to be more robust to this effect. DF, in contrast to the other integration methods, takes into account the agreement between sources through the DF Source Accuracy measure. This way, information shared between the sources over all windows is used to integrate data, instead of just using information from a single frame in MV, RND.

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 after around context windows sizes of 13 frames (or 1.3 seconds). For the other integration methods, we find that accuracy increases almost linearly with context window size. This shows that DF is better at finding useful shared knowledge between the microphones to make a good integration, compared to the other integration methods.

B. DF Source Accuracy.

An important part of data fusion is the computation of a DF source accuracy per source. DF source accuracy provides an agreement score for each source relative to the other sources, which is used for selecting the best values from the most accurate sources. This ranking can be used in CPS for the estimation of sensor quality. For example, in our application of data fusion for the integration of microphone classifications, DF source accuracy can provide a ranking of microphone trustworthiness, without having ground truth

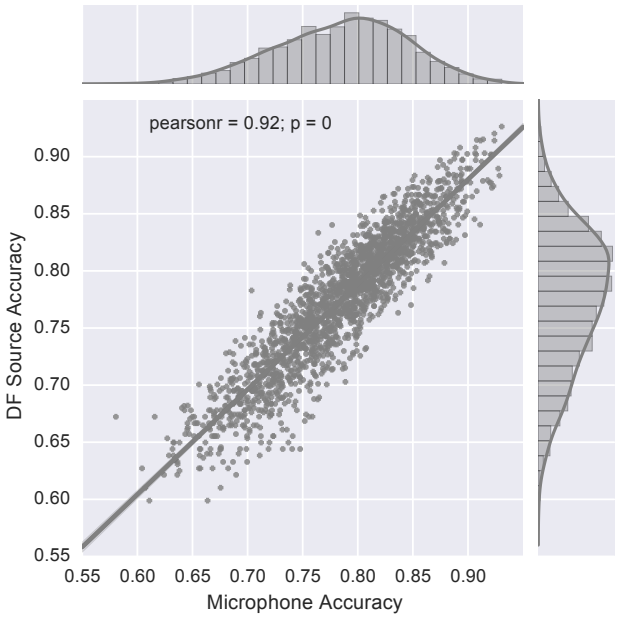

Fig. 6. Correlation between accuracy of each microphone and the DF source accuracy. The strong correlation shows that DF source accuracy is a strong indicator of the classification accuracy of the microphones.

knowledge. This knowledge can for example be used as side-channel information providing rotor or microphone quality. As an example, if a rotor is damaged, its sound will be different from the consensus and therefore will have a low DF source accuracy.

Microphone quality assessment. Investigating the relationship between DF source accuracy and the actual microphone classification accuracy provides insight whether data fusion is capable of assessing the relative quality of microphone state estimations. This relationship is shown in Fig. 6, in which microphone classification accuracy are plotted against the DF source accuracy. Fig. 6 shows the microphone accuracies and DF source accuracy for all microphones, window sizes, and cross-validation folds.

Fig. 6 shows that both DF source accuracies and the microphone accuracies the values follow a similar normal distribution with $\mu \approx 0.8$. Furthermore, it shows that the values are scattered along a diagonal line, indicating that a higher DF source accuracy is associated with a higher microphone accuracy, and vice versa. The strong correlation is confirmed by Pearson's measure of linear dependence. We find a Pearson's coefficient of 0.92 with a p-value of $p \ll 0.001$, indicating a strong linear correlation. These results show that using DF source accuracy, we can accurately assess the relative quality of microphone classifications without ground truth knowledge.

VI. Conclusions

We have shown that through audio content analysis and classification of quadcopter rotor sound, we can predict ascending, hovering and descending states of a quadcopter with accuracies over 94%. More specifically, we have shown that through data fusion of classifications from multiple microphones, we can improve ascending, hovering and descending state prediction compared to a single microphone. Furthermore, we have shown that we can accurately assess the relative quality of microphone classifications using data fusion source accuracy.

Our research contributes to a growing body of work of research into state prediction of Cyber-Physical Systems from the sound they make. Our results show the benefit of using multiple microphone side-channels to obtain state predictions with high assurance. Microphones are inexpensive sensors which are relatively hard to attack. Furthermore, the proposed feature extraction and machine learning prediction techniques are computationally cheap, yet robust. Therefore, we believe that microphones are an obvious side-channel choice for further research into obtaining state estimations with high assurance.

Acknowledgments

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] A. A. Cardenas, S. Amin, and S. Sastry, "Secure control: Towards survivable cyber-physical systems," in *Proceedings of the 28th International Conference on Distributed Computing Systems Workshops*. IEEE, 2008, pp. 495–500.

[2] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure state-estimation for dynamical systems under active adversaries," in *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2011, pp. 337–344.

[3] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems," in *Proceedings of the 3rd conference on Hot topics in security*. USENIX Association, 2008, p. 6.

[4] V. M. Igure, S. A. Laughter, and R. D. Williams, "Security issues in scada networks," *Computers & Security*, vol. 25, no. 7, pp. 498–506, 2006.

[5] Y. Chen, W. Xu, W. Trappe, and Y. Zhang, *Securing Emerging Wireless Systems: Lower-layer Approaches*. Springer Publishing Company, Incorporated, 2010.

[6] H. V. Koops and F. Franchetti, "An ensemble technique for estimating vehicle speed and gear position from acoustic data," in *Proceedings of the IEEE International Conference on Digital Signal Processing*. IEEE, 2015, pp. 422–426.

[7] M. Madain, A. Al-Mosaiden, and M. Al-khassaweneh, "Fault diagnosis in vehicle engines using sound recognition techniques," in *Proceedings of the IEEE International Conference on Electro/Information Technology*. IEEE, 2010, pp. 1–4.

[8] P. Henriquez, J. B. Alonso, M. A. Ferrer, and C. M. Travieso, "Review of automatic fault diagnosis systems using audio and vibration signals," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 5, pp. 642–652, 2014.

[9] X. L. Dong and D. Srivastava, "Big data integration," in *Proceedings of the IEEE 29th International Conference on Data Engineering*. IEEE, 2013, pp. 1245–1248.

[10] H. V. Koops, W. B. de Haas, D. Bountouridis, and A. Volk, "Integration and quality assessment of heterogeneous chord sequences using data fusion," in *Proceedings of the 17th International Society for Music Information Retrieval Conference*, 2016, pp. 178–184.

[11] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[12] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," in *Proceedings of the 17th Int. Society for Music Information Retrieval Conference*, 2016.

[13] T. K. Ho, "Random decision forests," in *Proceedings of the Third International Conference on Document Analysis and Recognition*, vol. 1. IEEE, 1995, pp. 278–282.

[14] ——, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[15] X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: the role of source dependence," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 550–561, 2009.

[16] X. L. Dong and D. Srivastava, "Big data integration," *Synthesis Lectures on Data Management*, vol. 7, no. 1, pp. 1–198, 2015.

[17] X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava, "Scaling up copy detection," in *Proceedings of the IEEE 31st International Conference on Data Engineering*. IEEE, 2015, pp. 89–100.

[18] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.