Quantifying the effects of varying light-visibility and noise-sound levels in practical multimodal speech and visual gesture (mSVG) interaction with aerobots

Ayodeji O. Abioye[1], Stephen D. Prior[1], Glyn T. Thomas[1], Peter Saddington[2], & Sarvapali D. Ramchurn[1].

[1]Faculty of Engineering & the Environment, University of Southampton, UK.
[2]Tekever Ltd, Southampton, UK.

Abstract

This paper discusses the research work conducted to quantify the effective range of lighting levels and ambient noise levels in order to inform the design and development of a multimodal speech and visual gesture (mSVG) control interface for the control of a UAV. Noise level variation from 55 dB to 85 dB is observed under control lab conditions to determine where speech commands for a UAV fails, and to consider why, and possibly suggest a solution around this. Similarly, lighting levels are varied within the control lab condition to determine a range of effective visibility levels. The limitation of this work and some further work from this were also presented.

Key words: mSVG (multimodal speech and visual gesture), aerobot, speech, nCA (navigation control autonomy)

Introduction

This paper is part of a series of research work investigating the use of novel HCIs in the control of small multirotor UAVs, with a particular focus on the multimodal speech and visual gesture (mSVG) interface [1]–[3]. The aim of this paper is to present the effect of varying visibility and noise levels in a practical multimodal speech and visual gesture (mSVG) control interaction with an aerial robot (aerobot) at a higher navigational control autonomy (nCA) level. Known limitations of the proposed system, from previous studies, suggests that 1) the mSVG method could be susceptible to speech corruption during capture, due to the noise generated by the multirotor propulsion systems and other loud ambient noise such as in stormy weathers, and 2) poor visibility levels could affect the visual gesture capture, as may be the case outdoor at night, or in cloudy or misty weather; although these effects were not quantified. Therefore, the extent of this limitation is being practically measured, in order to inform the possibility of developing techniques that could either extend the range of the mSVG method's usefulness or develop a way of working around it limitation.

The experiment study design would be discussed, and some experiment results presented in this paper. The study is being conducted on a computer-based UAV simulator, augmented with external hardware-in-the-loop components (single-board computers, cameras, microphones, speakers, and lighting systems), in order to interact with the physical world, and to provide the natural alternative method of mSVG interaction with a UAV operator. The effect of the varying noise levels and varying visibility levels on the mSVG interaction method is being measured by varying ambient noise level across five intervals between 50 dB and 90 dB. Similarly, the visibility level would be varied rom 10 Lux to around 5500 Lux.

Literature Review

A. multimodal speech and gesture interfaces

Multimodal speech and gestures interfaces are actively being developed for many mobile and stationary robotic systems. Ref [4] investigated multimodal speech and gesture communication with multiple UAVs in a search and rescue mission using the Myo armband device. The result of their simulation showed that a human operator could interact effectively and reliably with a UAV via multiple modalities of speech and gesture, in autonomous, mixed-initiative, or teleoperation mode [4]. Ref [5] investigated the use of natural user interfaces (NUIs) in the control of small UAVs, using a developed Aerostack software framework, combining several NUI methods and computer vision techniques. Their project was aimed at studying, implementing, and validating NUIs efficiency in human UAV interaction [5]. Ref [6], [7] investigated the performance of a speech and gesture multimodal interface for a soldier-robot team communication during an ISR mission. They also suggested the possibility of developing complex semantic navigation commands such as "perch over there (speech + pointing gesture), on the tank to the right of the stone monument (speech)" [6], [8]. Speech can be used to provide contextual information for the pointing gesture and vice versa. In a related research by [9], the researchers suggested that multimodal speech and gesture communication was a means to achieving an enhanced naturalistic communication, reducing workload, and improving the human-robot communication experience. Ref [10] also investigated the effectiveness of speech and gesture communication in soldier-robot interaction. Ref [11] investigated collocated interaction with flying robots, studying participants' behaviour around UAVs. Refs [12], [13] conducted elicitation studies to determine what gestures are considered intuitive for controlling UAVs. Ref [3] justifies the application of a multimodal speech and visual gesture interface for interacting with a patrol, search, and rescue UAV.

B. Speech interface

Speech is one component of the proposed mSVG aerobot control interface. In this method, controls are issued via voice commands. A microphone is used to detect the sound wave generated by an operator's voice commands, which is then convert to an electrical signal for processing. The operator's speech command may be identified by querying a database of speech command vocabulary with the captured speech signal,

for a match. Some popular audio speech recognition (ASR) toolkits are the Microsoft speech platform SDK, CMU PocketSphinx, and Googles web speech API [7].

In order to develop a speech control method for a UAV, one needs to take into account the average noise level generated by the UAV's multirotor propulsion system, in addition to the ambient noise levels. Refs [14], [15] both conducted an experiment to measure the noise level generated by small UAVs. In [15]'s experiments, five small UAVs were tested by flying the UAVs to a 1 m altitude, and placing a sound metre 1 metre adjacent to the UAV. The results obtained were as presented in Table 1. From these results, and for the purpose of this experiment, we can safely assume that the noise level generated by the small multirotor UAV is approximately 80 decibels. Sound reduces at a rate of 6 dB for every doubling of distance from a noise source [16]. Therefore, if a DJI phantom 2 is 75.8 dB at 1 m, then it would be 69.8 at 2 m, 81.8 dB at 0.5 m, and 87.8 dB at 0.25 m.

TABLE I
Noise levels generated by some small multirotor UAVs [15]

S/N	Small Multirotor UAV	Noise Levels (dB)
1	DJI Phantom 2	75.8
2	DJI Phantom 3 Pro	76.3
3	DJI Phantom 4 pro	76.9
4	DJI Inspire 2	79.8
5	Hover Cam	72.1

C. Gesture interface

The gesture interface is the second component of the mSVG interface. The method used in this work is similar to that described in [17] where hand gestures are recognised with the aid of convexity hall defects. A four-stage image processing operation of skin detection, noise elimination, convex hull and convexity defect processing with the aid of OpenCV algorithm libraries, to count the number of fingers being held up by a user.

Methodology

A. Experiment Design

As part of the larger research scope, human experiments were performed from which this paper's study was extracted. The experiment was conducted with the aid of a computer-based UAV simulator, augmented with external hardware-in-the-loop components (single-board computers, cameras, microphones, speakers, and lighting systems), in order to interact with the physical world, and to provide the natural alternative method of mSVG interaction with the UAV operator participant. The study participants were mostly sited in front of a three-screen UAV simulation computer workstation, during which the participant were asked to perform a series of task. The first task measured the effect of varying noise levels from 55 dB to 85 dB, generated from a Bose Sound link Mini speaker system, playing a pre-recorded multirotor UAV propeller-rotor noise. The second observed the effect of varying the ambient lighting conditions and how it affected a web cameras ability to capture finger gestures, against a white background as shown in Fig. 1, from 10 Lux to 1500 Lux.

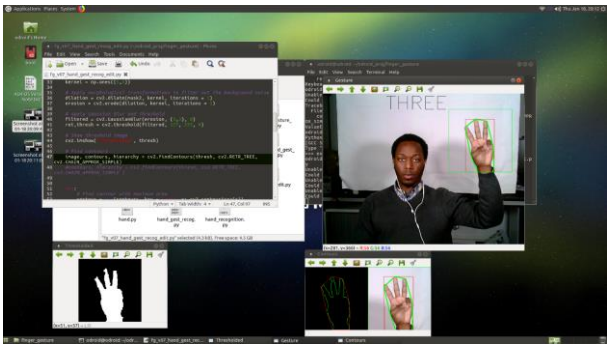

Fig. 1: Capturing finger gesture on a single board computer (SBC).

Both the speech and gesture control input are processed on an Odroid XU4 single board computer, with the aim of being able to plug this onto commercial UAV flight controller systems like the PixHawk, and have the SBC convert the high level nCA Tier I-III [2] or higher commands to lower levels that the flight controller can handle.

B. The experiments

Invited experiment participants are invited are issued with a series of 12 speech commands made from a vocabulary of 12 words. The commands are "go forward, go backward, step left, step right, hover, land, go forward half metre, go backward half metre, hover one metre, step left half metre, step right one metre, and stop" in that order. These commands are issued at quiet lab conditions of around 55 dB, and then repeated for 60 dB, 65 dB, 70 dB, 75 dB, 80 dB, and 85 dB noisy lab conditions. Successfully issued commands are recorded and have been presented in the result section of this paper.

The procedure is similar for the gesture capture, but with only five gesture commands, "one finger, two fingers, three fingers, four fingers, and five fingers" mapping to the following commands respectively, "forward, backward, right, left, and stop". The lab is made a dark as possible (the computer monitors generate some light), and the minimum lab lighting condition is measured, then two light variable LED light sources are used to generate various lighting intensity levels from between 10 Lux to 1500 Lux through two colour temperature of yellow (3200 K) and white (5600 K) lighting, as described in Fig. 2.

Fig. 2: Conducting gesture capture experiment under lab conditions

Results and Discussions

A. Speech

Fig. 3 is a series of bar charts showing the frequency of each vocabulary speech command word spoken (Purple), and the frequency of the successfully registered speech control command (Green). The Yellow coloured bar charts are the normalized ratios of the successfully registered speech control command (Green) to the spoken speech command (Purple). The numbers 1 to 12 on the x-axis represents the following single commands “go, forward, backward, right, left, step, hover, land, half, one, metre, and stop” respectively. For low noise levels of 55 dB and 65 dB, the normalized ration is mostly 1, except for a deep at command 7 (hover), were the system struggled to register some participants pronunciation of the word. These results were as collected from three participants.

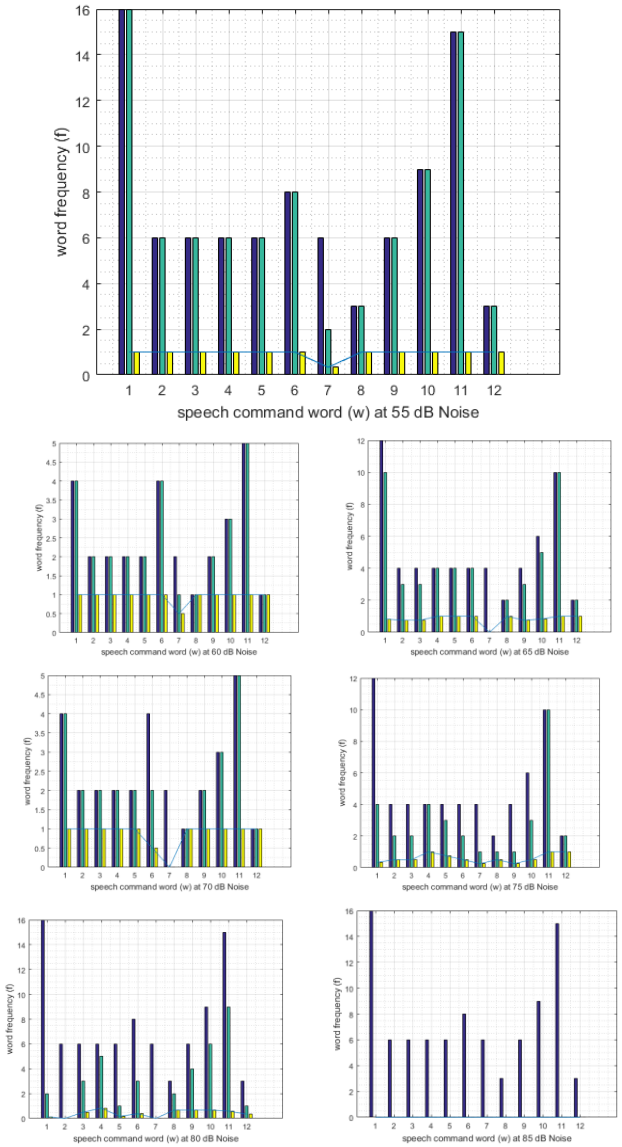

Fig. 3: mSVG UAV control speech level corruption with increasing ambient noise levels

Observe that the Yellow coloured bars keeps shrinking in size as the noise level increases from 65 dB to 80 dB, and at 85 dB, there is no Yellow coloured bar nor Green coloured bar either. This is because at 85 dB, the noise source completely drowns the microphone recording and no speech command could be registered. Reason for noise drowning speech at higher noise levels 85dB+ is because of the implementation technique of the ASR in which it determines the ambient noise level upon start up and begins recording at anything significantly higher than the ambient noise level. How significantly higher is adjustable by the microphone input recording gain using any available audio sound tool available on the SBC’s Linux installed software (Ubuntu Mint in this case). Perhaps an alternative method of cutting recording at higher noise levels could be implemented, but one wonders if this would be a feasible solution, because even after stopping the noise source to see if any speech was captured along with the noise, no speech could be deciphered, hence this method is also likely to fail. Two propositions that could work may be 1) having a model of the UAV rotor propulsion system noise and other likely environmental noise, and then subtract this from the captured recording via some software algorithm to determine if the speech signal could be identified. 2) the second option may be to have an array of microphone capture the ambient noise along with the speech, and if the direction from which the speech source is known, one could subtract the speech recordings of other microphones from the one facing the speech command giver, and perhaps the resultant audio input can be processed to determine the appropriate speech commands (a similar implementation as in the Amazon Echo devices).

B. Gesture

The results of the gesture capture showed that gesture could be successfully captured at low white (5600 K) light intensity from 36 Lux onwards, whereas yellow (3200 K) light intensity level of 850 Lux failed to register a gesture successfully. This could be because of the OpenCV background colour removal method used. Perhaps, upon adjusting some parameters, these may also work successfully. Also, the limit of 36 Lux presented here is probably the limit of the Odroid webcam used, perhaps, tweaking the gain could further improve the cameras ability to work in low light conditions, also there exists cameras designed to work effectively in low light conditions. However, one wonders if there a very low light condition may potentially affect the human operators ability to see a UAV and hence issued control commands.

Also, the Yellow light temperature did not affect white light capture, when both sources were combined at minimum conditions and at maximum conditions.

Conclusion

A. Summary

This paper presents and discusses a method of quantifying useful speech and gesture ranges for practical UAV applications. The rationale behind this thoughts were presented as a series of literature review on UAV related multimodal speech and gesture works currently being carried out. An experiment was designed and conducted to determine this limits, some results from the experiments were presented and discussed. It was observed that ambient noise level above 80 dB significantly affects speech capture. It was also observed that light intensity and the light colour temperature could affect gesture detection. Additionally, a number of factors were suggested that should be taken into account when designing or developing such a system.

B. Limitation

During the experiment, it was observed that ambient noise does not normally exist as a single dB value, but actually varies in pitch, frequency, and loudness with time and space. Hence visually average noise levels were used in experiment. A more optimised method could be considered for future investigations. Two attempts were allowed for participants to correct their speech commands, before recording observations, with the last command being recorded, similar to human-human interaction where one may have to repeat their words louder in noisy environment and the last command heard is executed.

C. Further works

As a further work from this, more participants would be recruited to confirm current observations and claims. Also, the next phase of this work would be to compare the mSVG and RC joystick in terms of training time, same nCA Tier task completion rate, and cognitive workload requirement on a ROS gazebo UAV simulator with hardware in the loop components.

Acknowledgement

This research was financially supported by the Petroleum Technology Development Fund (PTDF) of the Federal Government of Nigeria. Accessible via the following PTDF Reference Number: 16PHD052 and PTDF File Number: 862.

References

[1] A. O. Abioye, S. D. Prior, G. T. Thomas, and P. Saddington, "The Multimodal Edge of Human Aerobotic Interaction," in *International Conferences Interfaces and Human Computer Interaction*, 2016, pp. 243–248.

[2] A. O. Abioye, S. D. Prior, G. T. Thomas, P. Saddington, and S. D. Ramchurn, "Multimodal Human Aerobotic Interaction," in *Smart Technology Applications in Business Environments*, P. Isaías, Ed. IGI Global, 2017, pp. 39–62.

[3] A. O. Abioye, S. D. Prior, G. T. Thomas, P. Saddington, and S. D. Ramchurn, "The multimodal speech and visual gesture (mSVG) control model for a practical patrol, search, and rescue aerobot," in *19th Towards Autonomous Robotic Systems (TAROS) Conference - unpublished*, 2018, pp. x–x.

[4] J. Cacace, A. Finzi, and V. Lippiello, "Multimodal Interaction with Multiple Co-located Drones in Search and Rescue Missions," *CoRR*, vol. abs/1605.0, pp. 1–6, 2016.

[5] R. A. S. Fernandez, J. L. Sanchez-lopez, C. Sampedro, H. Bavle, M. Molina, and P. Campoy, "Natural User Interfaces for Human-Drone Multi-Modal Interaction," in *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2016, pp. 1013–1022.

[6] D. J. Barber, T. M. Howard, and M. R. Walter, "A multimodal interface for real-time soldier-robot teaming," vol. 9837, p. 98370M, 2016.

[7] J. Harris and D. Barber, "Speech and Gesture Interfaces for Squad Level Human Robot Teaming," in *Unmanned Systems Technology Xvi*, vol. 9084, R. E. Karlsen, D. W. Gage, C. M. Shoemaker, and G. R. Gerhart, Eds. SPIE, 2014.

[8] A. Borkowski and B. Siemiatkowska, "Towards semantic navigation in mobile robotics," *Graph Transform. Model. Eng.*, pp. 719–748, 2010.

[9] S. G. Hill, D. Barber, and A. W. Evans, "Achieving the Vision of Effective Soldier-Robot Teaming : Recent Work in Multimodal Communication," *Proc. Tenth Annu. ACM/IEEE Int. Conf. Human-Robot Interact. Ext. Abstr.*, pp. 177–178, 2015.

[10] R. K. Kattoju, D. J. Barber, J. Abich, and J. Harris, "Technological evaluation of gesture and speech interfaces for enabling dismounted soldier-robot dialogue," vol. 9837, p. 98370N, 2016.

[11] W. S. Ng and E. Sharlin, "Collocated interaction with flying robots," *Proc. - IEEE Int. Work. Robot Hum. Interact. Commun.*, pp. 143–149, 2011.

[12] J. R. Cauchard, L. E. Jane, K. Y. Zhai, and J. A. Landay, "Drone & me: an exploration into natural human-drone interaction," *Proc. 2015 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput.*, pp. 361–365, 2015.

[13] M. Obaid, F. Kistler, G. Kasparaviciute, A. E. Yantaç, and M. Fjeld, "HowWould You Gesture Navigate a Drone? A User-Centered Approach to Control a Drone," in *Proceedings of the 20th International Academic Mindtrek Conference*, 2016, pp. 113–121.

[14] R. Islam, S. Kelly, and A. Stimpson, "Small UAV Noise Analysis Design of Experiment," Duke University, 2016.

[15] T. Levin, "How loud is your drone? - The drone noise test of P2, P3P, P4P, and I2," *Online Webpage*, 2017. [Online]. Available: https://www.wetalkuav.com/dji-drone-noise-test/. [Accessed: 12-Oct-2017].

[16] R. A. Collman, "Is this too Noisy (or perhaps too Quiet)?," in *CIBSE London*, 2014, p. 96.

[17] S. Ganapathyraju, "Hand gesture recognition using convexity hull defects to control an industrial robot," in *2013 3rd International Conference on Instrumentation Control and Automation (ICA)*, 2013, pp. 63–67.