Speech Communication at the Presence of Unmanned Aerial Vehicles

Oliver Jokisch[1], Enrico Lösch[1] and Ingo Siegert[2]

[1] *Leipzig University of Telecommunications (HfTL), 04277 Leipzig, Germany, Email: jokisch@hft-leipzig.de*

[2] *Otto von Guericke University (OVGU), 39016 Magdeburg, Germany, Email: ingo.siegert@ovgu.de*

Introduction

Unmanned aerial vehicles (UAVs, aerial drones) find their way into different emerging application areas such as surveillance (disaster management or public safety), but also logistic tasks (urgent delivery of small goods) or collaborative robotic tasks (precision farming), widely relying on image processing, although sound is an important source of information too. A quite recent application of UAVs is the enforcement of a lockdown in several cities during the COVID-19 crisis [1, 2].

Due to rotor and other maneuver-related noise components, sound processing and, in particular, speech communication or speech signal processing at the presence of UAVs are challenging. Thus signal analysis methods have to consider both, various sound sources and dynamic room-acoustical settings, and the desired sound signals have to be separated from flight and wind noises. A deeper analysis of this issue is hindered by the fact that suitable reliable and reproducible recordings are missing. Therefore we compiled a database comprising speech data in combination with typical drone sounds (recorded during real flight maneuvers). Afterwards, we conducted objective speech quality measurements (simulating human perception) as well as speech recognition experiments with a state-of-the-art recognizer to get a first impression about the difficulties in human perception and automatic analyses that required various algorithmic and technical improvements.

In a next research step, we analyzed various speech and sound improvements utilizing on the one hand optimized microphone constellations and a portable 8-microphone array, and on the other hand algorithmic post-processing. Another problem approach, focusing on the construction of low-noise drones, was also discussed. Our contribution shortly describes the different acoustic setups and drone characteristics, potential methods to overcome the mentioned obstacles and the recent measurement results, followed by some conclusions.

Related Studies

Only a few research is reported on UAV-related acoustics. Hereby mostly the sound immission in humans – e.g. involving measurements of the sound pressure level [3, 4, 5] and spectral analyses of overflight noise [6] was in the focus. Some studies on influencing factors dealt with the number and type of rotor blades [7], the motor rotation speed [3] and the differences between quadcopters, tricopters or hexacopters [6]. Table 1 summarizes the average sound pressure levels (SPL) of popular UAVs at comparable measurement conditions [8].

While UAV-based image processing in civil and military environments was intensively studied, a targeted sound processing turns out to be challenging due to rotor and other noise at flying UAVs [6, 7]. Also the processing of environmental information was focused on electromagnetic signals or image processing, including object recognition with a variety of camera techniques. Except for the ultrasonic sensors, small consumer UAVs do not provide acoustic recording facilities. Nevertheless, possible applications of audio processing at the flying UAV include interesting use cases, such as the recognition of speech commands or the classification of environmental sounds.

Table 1: Average sound pressure level (SPL) of flying UAVs from [8] – the surveyed sample UAV is highlighted in gray.

UAV	Weight [kg]	Diam. [cm]	**SPL [dB]**	Ref.
DJI MavicAir	0.430	21.3	**98**	[9]
DJI MavicPro	0.734	33.5	**98**	[9]
DJI MavicProPlat.	0.734	33.5	**98**	[9]
Syma X5C	0.907	31.0	∼**82**	[3]
DJI Inspire1Pro	3.400	56.0	**81**	[10]
SwellPro Splashdr.	2.300	50.0	**80**	[10]
RC EyeOneXtreme	0.157	18.0	∼**66**	[3]
Quad-rotor MUAS	2.100	65.0	∼**64**	[5]

Consequently, the potential of sound or speech analysis directly at a UAV or in the near field was not systematically analyzed, although the additional acoustic or speech event analyses have some advantages over video-only analyses. Besides the lower transmission bandwidth of acoustic signals also the possibility of an intuitive interaction, as speech is the most natural way for humans to interact [11], has to be mentioned. For example, in the COVID-19 crisis where UAVs are used to enforce the lockdown, while at the same time reducing the contacts between humans, a bidirectional speech communication via the UAV would be helpful.

To systematically approach the audio characteristics of a flying UAV in a free field, we surveyed the blade passing frequencies (BPFs) and their harmonics in different flight maneuvers, including standard methods of noise suppression, in a first review [12], which proved the challenges of single-channel processing of wanted sounds and nearby speech commands. In [13], we tried to obtain more acoustic insights at the same, affixed UAV in a semi-anechoic chamber to ensure reproducible conditions, based on a two-channel microphone approach. We have generalized these preliminary results and discussed some UAV-based communication scenarios in [8].

The generally high SPL of UAVs is the main reason for the mentioned acoustic challenges since typical UAVs are originally designed for image capturing that requests a stable flight to reduce motion blur. Hence we suggested the design of specific low-noise UAVs to improve the signal-to-noise (SNR) ratio, e.g. by compromising on less-dynamic flight-stabilizing maneuvers, which are hardly relevant for audio capturing. In the last respective article [14] we focused on a constructive method to improve the signal capturing and analysis by using a lightweight 8-microphone array for beamforming, supplemented by a state-of-the-art method in post-filtering. In this study, we therefore summarize selected results of our acoustic experiments at or nearby unmanned aerial vehicles with a focus on speech signal processing.

UAV Test Setup and Speech Experiments

Baseline Setup

In [12], the sample quadcopter (DJI Mavic Pro [15]) with a weight of 734 g is supplemented by mounted equipment of 102 g comprising a recording smartphone and the omnidirectional micro Rode smartLav+ (frequency range 20 Hz to 20 kHz). To analyze the influence of the recording positions onto different flight maneuvers, use cases, and environments, the microphone was placed at different positions on the UAV. Acoustic measurements at a flying UAV pose some challenges, which affect the reducibility of the signal analysis and also the potential of noise filtering. The flight control together with micro-movements, varying rotor speeds and other dynamic factors, such as turbulent flow or reflections, can hardly be synchronized with harmonic or other analysis. All sounds were sampled at 44.1 kHz, 16 bit, WAV format (linear PCM) for five flight maneuvers including hovering, indoor and outdoor. Besides speech we also recorded environmental sounds from a car, motorcycle, and a church bell.

In our first speech experiment, drafted in Figure 1, the microphone-carrying UAV was hovering in a distance of 0.5 m to a loudspeaker that played random sequences of the German voice commands “Halt”, “Stopp”, “Start”, “Fliege”, “Eins”, “Zwei” and “Drei”, prerecorded from a male voice aged 22. To survey a potential signal enhancement by automatic noise reduction (ANR), we tested a single-channel ANR and a notch-filter method as well as low-pass filtering with a cut-off at 4 kHz.

The UAV-recorded command samples, including real-world noise and the noise-reduced versions of each command, were fed in random order to the Google Cloud Speech-to-Text API without additional training or adaptation to the specific noise conditions. In total, 735 command realizations were tested – on average 14 samples per command, including the original and up to four noise-reduced versions. Additionally, we ran an (instrumental) POLQA test to predict the overall speech quality as perceived by humans in an ITU-T P.800 absolute category rating (ACR) listening-only test [17, 18], which showed no significant enhancement in noise-reduced samples [8].

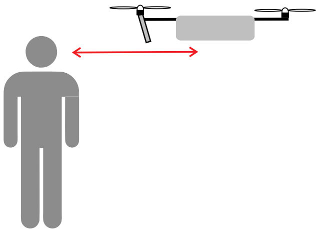

Figure 1: UAV interaction (stimuli by loudspeaker) [16].

Improved Setup by Beamforming

To potentially enable a significant SNR improvement, we re-designed the recording setup for the same quadcopter by using a lightweight microphone array for beamforming [14]. The audio signals are captured by the evaluation system “Distant Voice Acquisition Solution” (vicDIVA [19]), which includes an oval microphone array with eight MEMS microphones as shown in Figure 2. In conjunction with the vicSBM host system, the parameters directivity, main reception direction and signal amplification can be configured dynamically. The total weight of the mounted measuring system including cabling and battery amounts to 242 g.

All recordings were carried out in a quiet rural area under wind speeds of 0 km/h to 10 km/h. A loudspeaker (JBL Flip 4), located at 0.13 m above ground, played reproducible speech signals at a maximum volume. The speech samples were recorded simultaneously at the flying UAV under two conditions. The UAV is either directly hovering at 2 m height above the loudspeaker or flying in about 3.2 m ground distance (i.e. Euclidean distance of ca. 3.7 m) to the loudspeaker as depicted in Figure 3).

In addition to the UAV position, we also varied the two parameters directivity D and azimuth angle α. The values for D are 0 dB (bypass, omnidirectional), 18 dB,

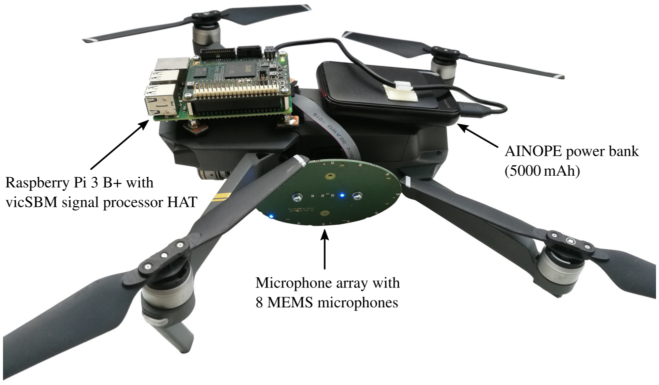

Figure 2: UAV-mounted audio system from [14].

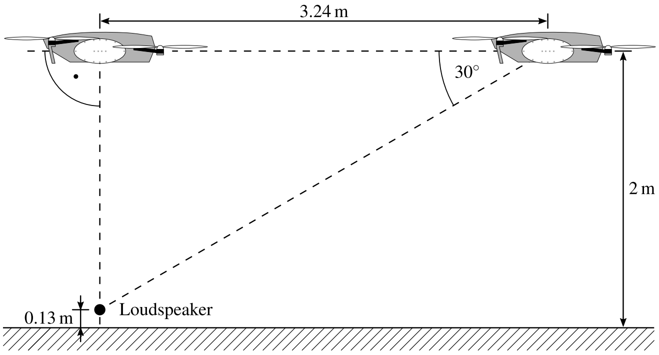

Figure 3: Audio recording setup from [14].

24 dB and 30 dB. As values for α 0° (highest sensitivity towards the front), 30°, 60°, and 90° (highest sensitivity towards the ground) were chosen. In total, 80 different settings were analyzed, comprising (a) the inherent noise of the UAV only (without voice signal), (b) the voice signal only (no UAV-induced noise) and (c) the voice under flight conditions (voice overlaid with the UAV's noise). For voice-signal recordings without UAV noise, the deactivated UAV was mechanically fixed at the respective position. As speech signal, the speech sequence "Male 1" from Appendix B.3.8 of ITU-T P.501 [20] was used. For easier localization of the speech sequences within the captured audio stream, the speech utterance was preceded with a pilot tone (5.5 kHz). The audio signals were recorded with a sampling rate of 16 kHz and a resolution of 24 bit. The vicSBM processes the individual microphone channels of the array and realizes beamforming and steering, at which the exact procedure is not published. The relevant speech chunks, needed for further processing, were manually extracted based on the pilot tones.

Results

Speech Recognition Test (Baseline Setup)

Table 2: Overall recognition rate (RR), rejections (Rej.) and RR* (RR without rejections) for 343 test signals from [12].

Noise reduction	SNR	Rej. %	RR %	RR* %
–	0	(100.0)	–	–
ANR	20 dB	**89.8**	10.2	100.0
Notch & LP	5 dB	69.4	28.6	93.3
Notch & ANR	25 dB	53.1	32.6	69.6
Notch	3 dB	46.9	**51.0**	**96.2**

The results of the command recognition are indicative and shall illustrate the potential for further studies only – the experimental details and more results are described in [16]. Expectedly, the harmonic components of both, speech and rotor sounds, overlap to a large extent. Even for a short speaker-microphone distance of 0.5 m, the signal-to-noise ratio (SNR) averages 0 dB only. Hence a command recognition without noise reduction is impossible, and a BPF-related filtering of the rotor harmonics obviously shows a limited success. The strongest noise reduction achieves an SNR improvement of about 20 dB but it still can not provide adequate input signals for the speech recognizer (rejection rate of 89.8 %). A notch-filtering with just 3 dB improvement seems to work in certain limits due to a targeted suppression of rotor harmonics. Regardless of a still unacceptable rejection rate of 46.9 %, the overall recognition rate (exclusive rejections) achieves 96.2 % after all.

Speech Enhancement Test (Improved Setup)

At first, the effects of the signal processing automatically executed on the vicDIVA/vicSBM hardware, namely beamforming and beam steering, are examined. The power spectrum shows the typical noise spectrum of the DJI Mavic Pro, which is already described in [12]. The intended directional effect is only achieved above 3 kHz as the attenuation of the UAV noise employing beamforming is frequency-dependent. Also, changing the azimuth angle (beam steering) for the selected values does not show any effect on the attenuation of the UAV noise in the near field condition. The highest signal level can be observed for the case that the main beam is aligned with the source (α=30°), which is the expected behavior. Furthermore, the low-pass behavior of the signal processing is visible, for the condition that sound waves are incoming from the side. To estimate the optimal SNR, the spectra of both, the UAV noise and the speech signal, are compared for two settings in Figure 4. It is apparent that even with the highest directivity of 30 dB, the interference power is higher than the power of the wanted speech signal, especially in the important frequency range of human speech (< 3 kHz).

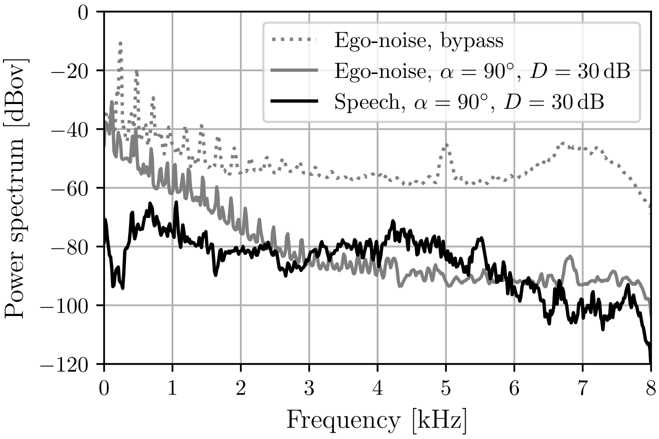

Figure 4: Target speech vs. UAV noise (best case) [14].

To additionally improve the SNR, advanced techniques such as Quantile Based Noise Estimation (QBNE) or Adaptive Quantile Based Noise Estimation (AQBNE) were examined concerning their suitability. QBNE estimates the noise during speech and non-speech parts without the use of a voice activity detector [21]. AQBNE further extends QBNE by using independently determined q-values for each frequency instead of a predefined, constant q-value as used in QBNE [22]. Details of our implementation for UAV-noise estimation and subtraction are described in [14]. We could not achieve sufficient noise reductions with either of both methods, which confirmed a subjective auditive perception: Though the resulting signal power is significantly reduced, there is no improvement in the speech intelligibility.

Discussion and Conclusions

Our contribution surveyed different approaches to enable speech communication at UAVs. Hereby, it was first analyzed to which extend state-of-the-art recognition systems are able to process speech recorded under UAV noise. Afterwards baseline and more advanced techniques for SNR improvements were presented.

Although the noise reduction for the baseline setup reaches an SNR improvement of about 20 dB, an adequate speech recognition is still not possible due to the high rejection rates. By the described beamforming and steering method, the SNR could be enhanced – in the best test cases by about 35 dB for signals above 3 kHz.

In the optimal case, the main beam must be aligned to the signal source with maximum directivity. Nevertheless, the noise attenuation significantly decreases for values below 3 kHz, hindering a proper speech processing. Compared to an omnidirectional micro setup, an SNR improvement of only 10 dB can be expected around 1 kHz, which is also reflected by the perceptive impression of typical speech samples. A consecutive noise reduction does not result in a significant signal enhancement either. In the context of heavily interfered, wanted speech signals at the UAV, a stable distinction between speech and noise is unlikely for most of the application scenarios.

We will optimize the measurement setup by an advanced microphone array for distinctive signal separation including the processing of separate microphone signals to support multichannel noise estimation. We will also analyze concrete UAV control data, e.g. the engine speeds, to enable a parameter adaptation during the noise filtering.

Acknowledgment

This work was supported by the EU project "Collaborative strategies of heterogeneous robot activity at solving agriculture missions controlled via intuitive human-robot interfaces (HARMONIC)" within the "ERA.Net RUS Plus/Robotics" program 2018–2020 (project ID 99) co-funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01 DJ18011 and the Russian Foundation for Basic Research under Grant No. 18-58-76001_ERA.Net.

References

[1] M. Bourdon and R. Moynihan. One of the largest cities in france is using drones to enforce the country's lockdown after the mayor worried residents weren't taking containment measures seriously. *Business Insider*, March 2020.

[2] R. Estrada and M. Arturo. The uses of drones in case of massive epidemics contagious diseases relief humanitarian aid: Wuhan-covid-19 crisis. *SSRN Electronic Journal*, February 2020.

[3] U. Papa, G. Iannace, G. Del Core, and G. Giordano. Determination of sound power levels of a small uas during flight operations. In *Proc. INTER-NOISE*, volume 45, pages 216–226, Hamburg, August 2016.

[4] M. Miesikowska. Analysis of signal of x8 unmanned aerial vehicle. In *Proc. IEEE Conf. Signal Processing: Algorithms, Architectures, Arrangements & Applications*, pages 69–72, Poznan, Poland, September 2017.

[5] N. Kloet, S. Watkins, and R. Clothier. Acoustic signature measurement of small multi-rotor unmanned aircraft systems. In *International Journal of Micro Air Vehicle*, volume 9(1), 7, pages 3–14, February 2017.

[6] R. Cabell, F. Grosveld, and R. McSwain. Measured noise from small unmanned aerial vehicles. In *Proc. INTER-NOISE/NOISE-CON*, volume 252, pages 345–354, Hamburg, June 2016.

[7] N. Intaratep, W. Alexander, W. Devenport, S. Grace, and A. Dropkin. Experimental study of quadcopter acoustics and performance at static thrust conditions. In *Proc. 22nd Aeroacoustics Conference (AIAA/CEAS)*, Lyon, France, June 2016.

[8] O. Jokisch, I. Siegert, M. Maruschke, T. Strutz, and A. Ronzhin. Don't talk to noisy drones - acoustic interaction with unmanned aerial vehicles. In *Proc. 21th Intern. Conference on Speech and Computer (SPECOM). In: Springer LNAI*, volume 11658, pages 180–190, Istanbul, Turkey, August 2019.

[9] D. Miljkovic. Methods for attenuation of unmanned aerial vehicle noise. In *41st Intern. Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 914–919, May 2018.

[10] F. Christiansen, L. Rojano-Donate, P. T. Madsen, and L. Bejder. Noise levels of multi-rotor unmanned aerial vehicles with implications for potential underwater impacts on marine mammals. *Front. Mar. Sci.*, 26, 2016.

[11] J. M. Carroll. Human computer interaction - brief intro. In M. Soegaard and R. F. Dam, editors, *The Encyclopedia of Human-Computer Interaction.* The Interaction Design Foundation, Aarhus, Denmark, 2 edition, 2013.

[12] O. Jokisch and D. Fischer. Drone sounds and environmental signals – a first review. In P. Birkholz and S. Stone, editors, *Proc. 30th ESSV Conference (Studientexte zur Sprachkommunikation, vol. 93)*, pages 212–220, Dresden, March 2019.

[13] O. Jokisch. A pilot study on the acoustic signal processing at a small aerial drone. In *Proc. 14th International Conference on Electromechanics and Robotics "Zavalishin's Readings" ER(ZR). In: Springer Smart Innovation, Systems and Technologies*, volume 154, chapt. 25, pages 305–317, Kursk, Russia, April 2019.

[14] E. Lösch, O. Jokisch, A. Leipnitz, and I. Siegert. Reduction of aircraft noise in uav-based speech signal recordings by quantile based noise estimation. In A. Wendemuth, R. Boeck, and I. Siegert, editors, *Proc. ESSV Conference (Studientexte zur Sprachkommunikation, vol. 95)*, pages 149–156, Magdeburg, March 2020.

[15] DJI Mavic Pro, 2018. Retrieved on 28 January 2020. URL: www.dji.com/en/mavic.

[16] D. Fischer. *Untersuchung von Geräusch- und Sprachsignalen beim Einsatz von Flugdrohnen (UAV) [in German].* Bachelor's thesis, HfT Leipzig, November 2018.

[17] ITU-T. Methods for objective and subjective assessment of speech quality (POLQA): Perceptual Objective Listening Quality Assessment. REC P.863, ITU, 2014. URL: www.itu.int/rec/T-REC-P.863-201409-I/en.

[18] ITU-T. Methods for subjective determination of transmission quality. Recommendation P.800, ITU, 1996. URL: www.itu.int/rec/T-REC-P.800-199608-I/en.

[19] A development kit for the distant voice acquisition (vicDIVA Kit), June 2019. Retrieved on 28 January 2020. URL: www.voiceinterconnect.de/en/sdk_beamforming.

[20] ITU-T. Test signals for use in telephonometry, Recommendation P.501, March 2017. URL: www.itu.int/rec/T-REC-P.501-201703-I/en.

[21] A. Fischer V. Stahl and R. Bippus. Quantile based noise estimation for spectral subtraction and wiener filtering. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1875–1878, June 2000.

[22] C.S. Bonde, C. Graversen, A.G. Gregersen, K.H. Ngo, K. Nørmark, M. Purup, T. Thorsen, and B. Lindberg. Noise robust automatic speech recognition with adaptive quantile based noise estimation and speech band emphasizing filter bank. In *Nonlinear Analyses and Algorithms for Speech Processing. NOLISP2005*, volume 3817, pages 291–302, 2006.