Reduction of Aircraft Noise in UAV-Based Speech Signal Recordings by Quantile Based Noise Estimation

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

[1]*Institute of Communications Engineering, HfT Leipzig, Germany*
[2]*Institute for Information Technology and Communications, OvG University Magdeburg*
{enrico.loesch, oliver.jokisch, alexander.leipnitz}@hft-leipzig.de, siegert@ovgu.de

Abstract: In this article we survey, whether the signal-to-noise ratio of speech signals, superimposed by flight noise, can be significantly improved by advanced noise-estimation methods, such as Quantile Based Noise Estimation (QBNE) or Adaptive Quantile Based Noise Estimation (AQBNE) and a spectral subtraction of the respective noise components. Our test object, a typical commercial UAV (DJI Mavic Pro), has been extended by a lightweight 8-microphone array (vicDIVA). The speech recordings in a free-field environment are processed with a battery-powered Raspberry Pi 3, attached to the UAV. The source of the speech signals is located on the ground. During our recordings, the UAV is hovering over the sound source. In addition to the noise-estimation methods QBNE and AQBNE, different distances between sound source and UAV as well as the directional effects of the microphone array (beamforming and steering) are investigated.

1 Introduction

In the last decade, the civilian operation areas of unmanned aerial vehicles (UAV) grew steadily, including surveillance tasks, the inspection of industrial structures, monitoring tasks in agriculture, data collection tasks in science, as well as the rescue of persons during disasters [1]. Other scenarios, such as UAVs for parcel delivery or autonomous flight taxis are currently under development. Applications that record and analyze acoustic signals with a rotary-wing UAV are still experimental and challenging, mainly due to the strong, non-stationary noise generated by such an aircraft. Acoustic or speech event analyses directly at UAVs have some advantages over video-only analyses, including a lower transmission bandwidth of acoustic signals in comparison to video signals [2] as well as the possibility of intuitive interaction [3].

So far sound or speech analysis with UAVs was just rarely investigated: A typical audio application is the recording of acoustic UAV signatures by other surveillance UAVs [4], sound immission in humans, involving measurements of the sound pressure level [5] and spectral analyses of overflight noise [6]. Our current contribution is focusing on the improvement of the signal-to-noise ratio in speech signals by applying more advanced methods of noise analysis such as Quantile Based Noise Estimation (QBNE) or Adaptive Quantile Based Noise Estimation (AQBNE) and the subsequent reduction of the noise components as demonstrated in [7, 8].

For this purpose, we recorded speech samples directly at a flying UAV (DJI Mavic Pro [9]). Our previous, first review of the audio characteristics for this UAV type in the free field showed dominating and significantly varying blade passing frequencies (BPFs) and associated harmonic components depending on the recording position and the flight maneuver [10]. The single-channel analysis of environmental sounds or even nearby speech commands, involving standard methods of noise suppression, turned out to be challenging. In [11], 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, with limited success. In a next step [12], we generalized our preliminary results and discussed some UAV-based communication scenarios and challenges more generic. We suggested, amongst others, the design of specific 'low-noise' UAVs to improve the signal-to-noise ratio, e.g. by compromising on less-dynamic flight-stabilizing maneuvers, which are hardly relevant for audio-oriented versus video tasks (that usually require a higher flight stability).

In this contribution, we focus on a constructive method to improve the signal capturing and analysis, by using a lightweight microphone array for beamforming, supplemented by state-of-the-art methods in post-filtering.

2 Experimental methods and data

2.1 Measurement system and recording scenario

The audio signals are captured by the evaluation system "Distant Voice Acquisition Solution" (vicDIVA [13]) from the company voice INTER connect GmbH. It consists of the hardware module vicSBM for signal processing and an oval microphone array with eight MEMS microphones. The system was developed for far field voice recordings. The vicSBM hardware module is designed as a "Hardware Attached on Top" (HAT) solution for the host platform Raspberry Pi 3. With the host system, the directivity, the main reception direction and the signal amplification can be dynamically configured, and the signal output from vicSBM can be recorded or processed. The measuring system, including cabling and battery, has a total weight of 242 g.

Figure 1 illustrates the test object DJI Mavic Pro with the fully-installed measuring system. Attention was paid to a balanced weight distribution. The UAV itself has sensors on the front and bottom for a stable flight position. In addition, there are air outlets on the back for actively cooling the electronic components. To avoid mutual interference between the measuring system and the UAV, the microphone array was placed on the left side of the UAV between the rotors.

All recordings were carried out in a rural area, with a distance of ≈ 600 m to the next road and ≈ 800 m to the next town. During the measurements, the temperature on the ground ranged between 2° C and 12° C, the relative humidity between 64 % and 77 % and the wind speed from 0 km/h to 10 km/h. A loudspeaker (JBL Flip 4), located at 0.13 m over ground, played reproducible speech signals at a maximum volume that were simultaneously recorded at the flying UAV directly hovering at 2 m height above the loudspeaker or in 3.24 m ground distance (i.e. Euclidean distance of 3.74 m) to the loudspeaker (see Figure 2). In addition to the UAV

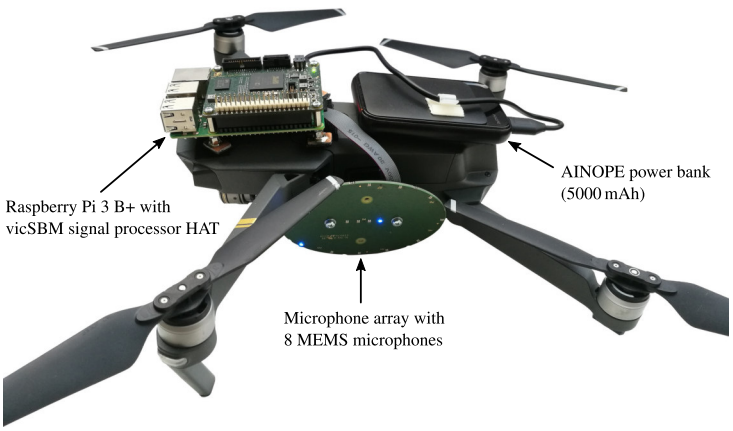


Figure 1 – Audio measurement system vicDIVA [13] mounted on the test UAV (DJI Mavic Pro [9])

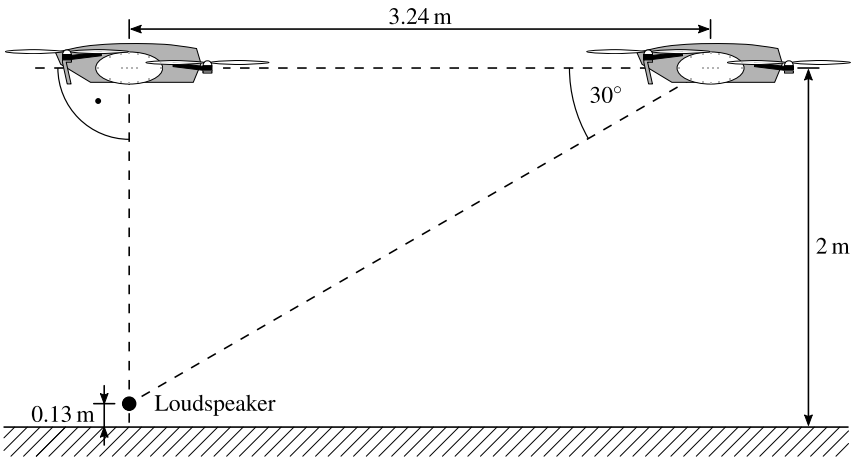
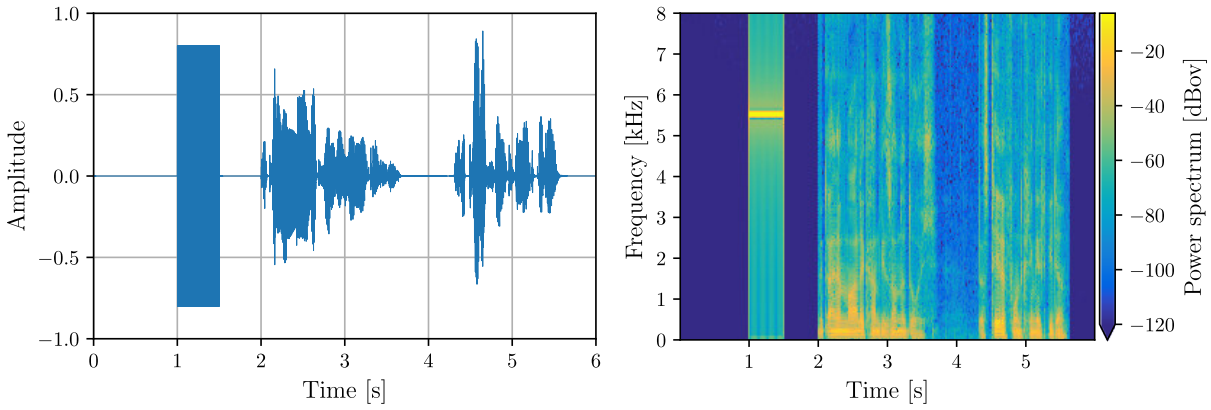


Figure 2 – Schematic representation of the combined UAV/audio measurement setup

position, we varied the directivity D step-wise from 0 dB (bypass, omnidirectional) via 18 dB and 24 dB to 30 dB, as well as the azimuth angle α from 0° via 30° and 60° to 90°. At $\alpha = 0°$ the microphone array has the highest sensitivity towards the front of the UAV and at $\alpha = 90°$ towards the ground with a horizontal flight attitude. The inherent noise of the UAV (without voice signal), the voice signal (without UAV-induced noise) and the voice recording during the flight (voice overlaid with the noise of the UAV) were examined separately in the settings mentioned above. For voice-signal recordings without UAV noise, the deactivated UAV was mechanically fixed at the respective position. A total of 80 different settings were analyzed.

2.2 Audio recording and signal preprocessing

We used the speech sequence "Male 1" according to Appendix B.3.8 of ITU-T P.501 [14] for all measurements. To localize the speech sequences within the captured audio stream, we preceded the utterance of the male speaker by a pilot tone with a frequency of 5.5 kHz. Figure 3 visualizes a typical test sequence in the time and spectral domain[1]. In this example, the speech utterance begins 0.5 s after the end of the pilot tone and lasts for about 3.5 s.



(a) Test signal in the time domain

(b) Spectrogram of the test signal

Figure 3 – Typical test signal – pilot tone at 5.5 kHz, followed by the utterance of a male speaker

The audio signals were recorded with the measuring system at a sampling rate of 16 kHz with a resolution of 24 bit. The vicSBM processes the individual microphone channels of the array and realizes the beamforming and steering, at which the exact procedure is not published.

[1]The presented spectrograms are based on 1,000 data points per segment, Hamming window and 50 % overlap.

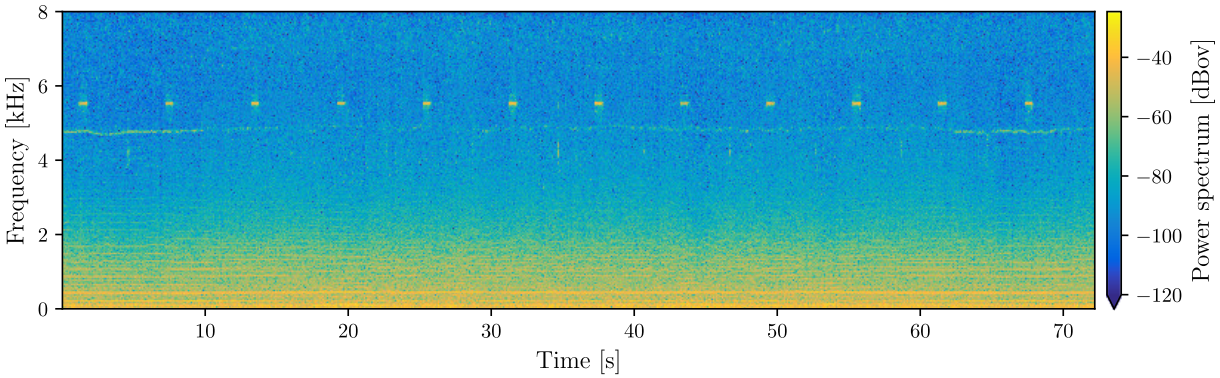


Figure 4 – Spectrogram of a flight example, including a sequence of 11 speech utterances. (The UAV hovers over the signal source with $\alpha = 90°$ and $D = 30\,\text{dB}$.)

In order to take account of random influences (e.g. wind or unstable flight positions), each measurement was repeated ten times at least. Figure 4 displays a sample sequence of a UAV flight including the voice recording of 11 complete speech utterances. The start and the end of the speech intervals can be determined by means of the distribution of pilot tones in the spectrogram. Only the speech chunks (e.g. second 2 to 6 in Figure 4) are processed further.

2.3 Quantile Based Noise Estimation (QBNE)

We assume that the power spectrum of the observed signal $X(f)$ is the result of a superposition of the power spectrum of the speech signal $S(f)$ with the power spectrum of the interference signal $N(f)$. Therefore, the following applies:

$$X(f) = S(f) + N(f). \tag{1}$$

If $N(f)$ was known, the power spectrum of the undisturbed speech signal would be determined with $S(f) = X(f) - N(f)$. Therefore, the best possible estimate of the noise component is a central task in noise reduction. The advantage of the QBNE method described in [7] is that no detection of the speech pause is required for the estimation, since the speech signal does not constantly occupy the entire spectrum. We apply the procedure as follows: The observed signal $x(t)$ is divided into I frames, which overlap by 50 %. After applying a Hamming window, the performance spectra are calculated and then sorted by frequency and size. The procedure results in a sequence for the mth frequency: $(X_0(f_m), X_1(f_m), \ldots, X_i(f_m), \ldots, X_{I-1}(f_m))$, in which $X_{I-1}(f_m)$ is the highest observed power $X_{\max}(f_m)$. The qth quantile of the sequence represents the estimate of the interference power $\tilde{N}(f_m)$ of the mth frequency:

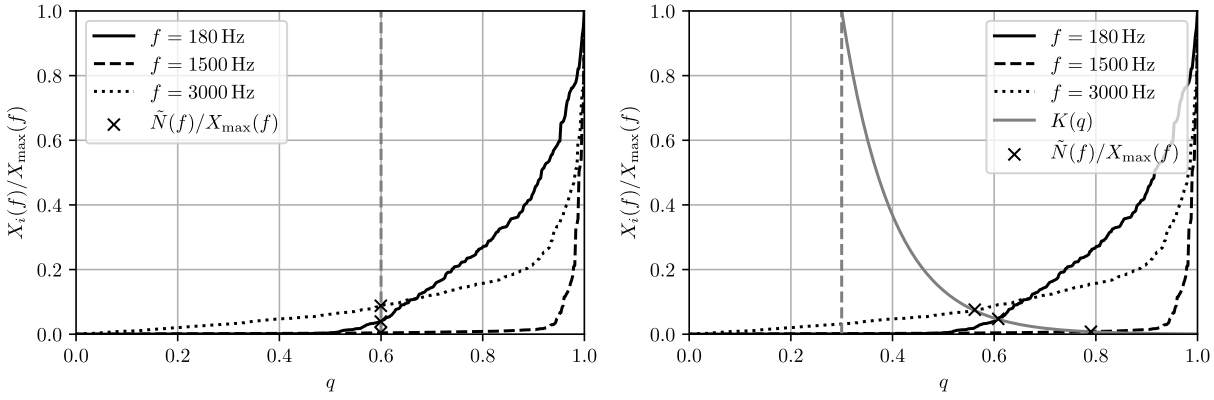
$$\tilde{N}(f_m) = X_{\lfloor q(I-1) \rfloor}(f_m). \tag{2}$$

For example, the medians are obtained with $q = 0.5$, the Figure 5a represents the normalized quantile of a disturbed speech signal for three frequencies and the estimate of the interference power for $q = 0.6$.

2.4 Adaptive Quantile Based Noise Estimation (AQBNE)

With the QBNE method, q is identical for all frequencies and independent of the observed powers of the respective frequency. In [8], an extension of the QBNE method for the adjustment of q is presented. It is assumed that at strongly disturbed frequencies, the interference energy is higher during the majority of the observation period and thus smaller q values lead to better overall results. The interference power is approximated using the estimation function:

$$K(q) = e^{(q_{\min} - q)\tau}, \tag{3}$$



(a) QBNE **(b)** AQBNE

Figure 5 – Normalized quantile of a noisy speech sample in QBNE and AQBNE at different frequencies

at which $q_{\min}$ is the smallest permissible value for q, and τ determines the slope of the curve. As depicted in Figure 5b, the intersections of $K(q)$ with the quantile curves are the estimated values $\tilde{N}(f)$.

2.5 Spectral subtraction

After estimating the noise by the previously described methods QBNE or AQBNE, a part of the inherent noise components can be removed in the regarding frequency range.
With the spectral subtraction:

$$\tilde{S}(f_m) = \begin{cases} X(f_m) - \tilde{N}(f_m) & \text{if } X(f_m) > \tilde{N}(f_m) \\ 0 & \text{else} \end{cases}, \tag{4}$$

the power spectrum of the speech signal is estimated as: $\tilde{S}(f_m)$. After an inverse FFT, the according estimate in the time domain is:

$$\tilde{s}(t) = \mathrm{FFT}^{-1}\left\{\sqrt{\tilde{S}(f)}\, e^{\mathrm{j}\Phi}\right\}, \tag{5}$$

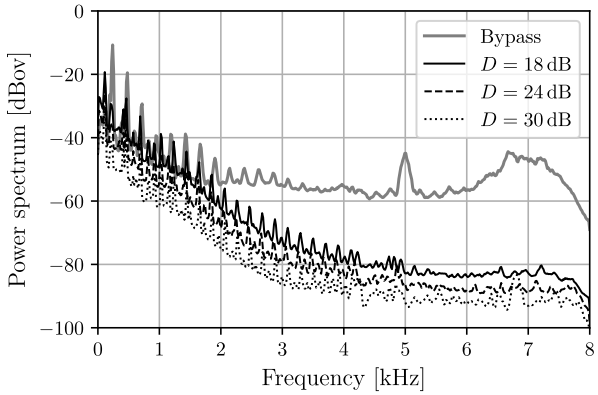
at which the phase Φ of the observed spectrum is supplemented.

3 Preliminary results and discussion

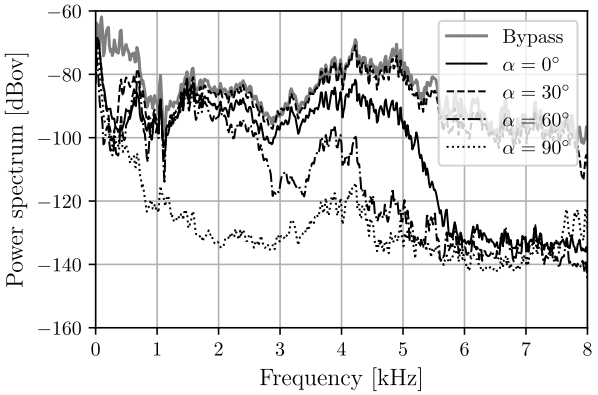
3.1 Beamforming and steering

Figure 6a depicts the power spectrum of the observed UAV noise, depending on the directivity (beamforming). The gray curve recorded with omnidirectional microphone settings shows the typical noise spectrum of the DJI Mavic Pro, which is already described in [10]. It can be clearly seen that the attenuation of the UAV noise by means of beamforming is frequency-dependent and that the intended directional effect is only achieved above 3 kHz. From a threshold of about 4 kHz, the noise can be damped by around 35 dB. The change in the azimuth angle (beamsteering) in the area $\alpha = \{0°, 30°, 60°, 90°\}$ demonstrates at $D = 30\,\mathrm{dB}$ no effects on the attenuation of the UAV noise in the near field.

Figure 6b depicts a power spectrum of the speech sequence at $D = 30\,\mathrm{dB}$ depending on α for the case that the UAV is close to the signal source. As expected, the signal level is the highest, if the main beam is aligned with the source (α=30°). In addition, the low-pass behavior of the vicSBM signal processing is visible, if the waves are falling in from the side.

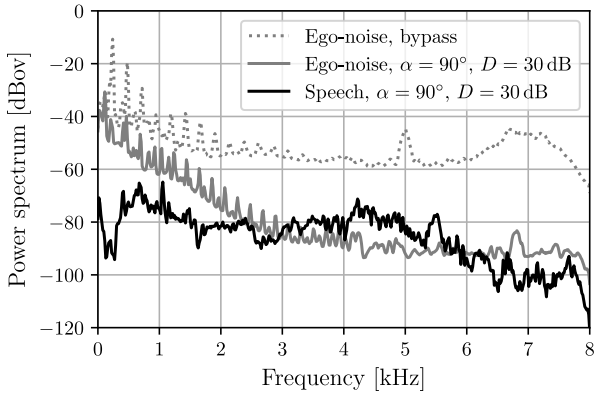


(a) Power of UAV ego-noise, depending on the directivity (beamforming), $\alpha = 90°$

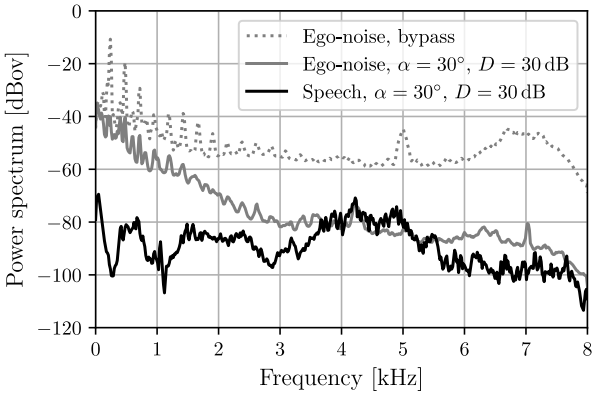


(b) Power of clean speech next to source, depending on azimuth (beamsteering), $D = 30\,\mathrm{dB}$

Figure 6 – Measurement examples: **(a)** UAV ego-noise (w/o speech) vs. **(b)** speech (w/o UAV activity)



(a) UAV hovers over the signal source.



(b) UAV hovers beside the signal source.

Figure 7 – Comparison of the interference power (UAV noise) and the power of the speech signal

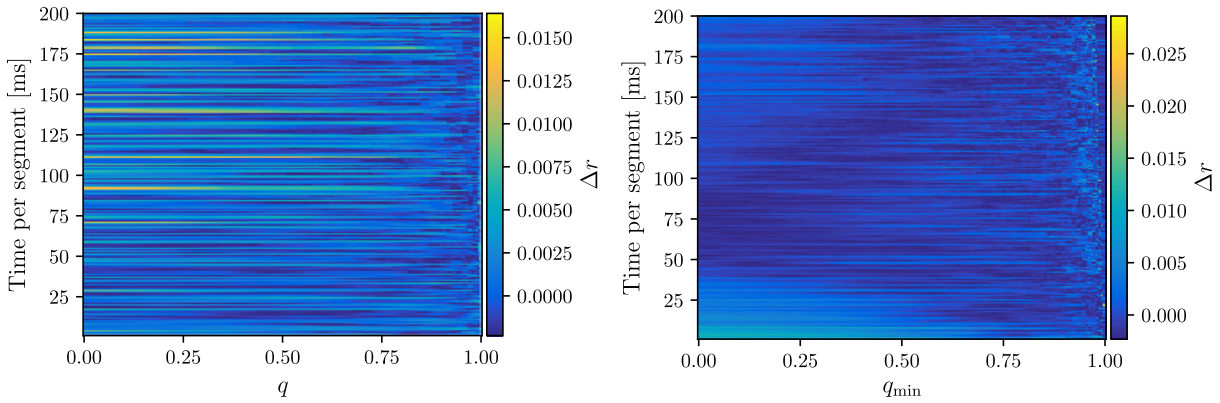
To estimate the signal-to-noise ratio (SNR) in the best case, the spectra of the UAV noise and the speech signal are compared for both positions as demonstrated in Figure 7. It is obvious that even with a directivity of 30 dB, especially in the important frequency range of human speech ($< 3\,\mathrm{kHz}$), the interference power is higher than the power of the wanted speech signal. Therefore, further techniques are required to improve the SNR.

3.2 Additional noise reduction by QBNE or AQBNE

To improve the SNR, QBNE and AQBNE were applied to the output signal of the vicSBM processing as described. Only the most favorable case will be considered below, i.e. the UAV is hovering over the signal source, and the recording takes place with a maximal directivity and alignment of the main beam to the source, see Figure 7a. In our tests, the segment length, q or $q_{\min}$ and τ are varied in the range from 2 to 15.

The results are evaluated based on the Pearson's correlation. The correlation between the signal prior and after the noise reduction with a recording of the same speech sequence without UAV activity was calculated, and the difference Δr of the amounts was determined. The results are visualized in Figure 8. The improvement of the correlation coefficient reaches the maximums of 0.016 with QBNE and 0.029 with AQBNE.

Thus, no significant noise reduction could be achieved with either method, which confirms our subjective auditive perception: Although the resulting signal power is significantly reduced, there is no improvement in the speech intelligibility.



(a) QBNE **(b)** Adaptive QBNE ($\tau = 6$)

Figure 8 – Variation of the correlation coefficient depending on the method parametrization

4 Conclusions

In this contribution we analyzed, whether a combined system of a microphone array and its processing hardware (evaluation kit vicDIVA) mounted on a test UAV (DJI Mavic Pro) can capture and analyze speech during a flight operation with a suitable quality. With the limited possibilities of beamforming and steering in the evaluation kit, the SNR can be improved in the best case (alignment of the main beam to the signal source and maximum directivity) by about 35 dB for signals above 3 kHz. The attenuation of the UAV noise significantly decreases below 3 kHz. For example at 1 kHz, an SNR improvement of only 10 dB can be expected, compared to an omnidirectional microphone. Despite beamforming and steering, it is not possible to well-understand the speech utterances, superimposed by the UAV noises in the recordings. The application of the methods QBNE or AQBNE for the subsequent noise reduction does not show a relevant signal improvement. By the described kind of noise estimation and reduction, a following distinction between speech and noise is not possible in the context of the heavily affected, wanted signal frequencies in our experiment.

For the further work, the measurement setup should be improved with a quieter UAV and a more appropriate microphone array, which allows a higher signal separation. Additionally, the processing of the individual microphone signals would be advantageous to test multi-channel noise estimation methods. It can also be assumed that the concrete flight data provided by the UAV control – especially the engine speeds – can improve a decided parameter adaptation of the noise filtering method.

The lessons learned so far will be decidedly discussed in [15] by a comparison of selected results from our previous studies [10, 11, 12] and the current contribution.

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] HASSANALIAN, M. and A. ABDELKEFI: *Classifications, applications, and design challenges of drones: A review*. *Progress in Aerospace Sciences*, vol. 91, p. 99–131, May 2017.

[2] SIEGERT, I., A. F. LOTZ, L. L. DUONG, and A. WENDEMUTH: *Measuring the impact of audio compression on the spectral quality of speech data*. In *Elektronische Sprachsignalverarbeitung*, pp. 229–236. 2016.

[3] CARROLL, J. M.: *Human computer interaction - brief intro*. In M. SOEGAARD and R. F. DAM (eds.), *The Encyclopedia of Human-Computer Interaction*, p. s.p. The Interaction Design Foundation, Aarhus, Denmark, 2 edn., 2013.

[4] RASCON, C., O. RUIZ-ESPITIA, and J. MARTINEZ-CARRANZA: *On the use of the aira-uas corpus to evaluate audio processing algorithms in unmanned aerial systems*. *Sensors*, 19(18), pp. 3902–3921, September 2019.

[5] MIESIKOWSKA, M.: *Analysis of signal of x8 unmanned aerial vehicle*. In *Proc. IEEE Conf. Signal Processing: Algorithms, Architectures, Arrangements & Applications (SPA)*, pp. 69–72. Sept. 2017.

[6] CABELL, R., F. GROSVELD, and R. MCSWAIN: *Measured noise from small unmanned aerial vehicles*. In *Proc. INTER-NOISE/NOISE-CON*, vol. 252, pp. 345–354. June 2016.

[7] V. STAHL, A. F. and R. BIPPUS: *Quantile based noise estimation for spectral subtraction and wiener filtering*. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, vol. 3, p. 1875–1878. June 2000.

[8] BONDE, C. S., 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, Barcelona, Spain, April 2005. Lecture Notes in Computer Science, Springer*, vol. 3817, p. 291–302. 2006.

[9] DJI Technology Ltd.: *DJI Mavic Pro*. 2018. URL `www.dji.com/en/mavic`. Retrieved 28/01/2020.

[10] JOKISCH, O. and D. FISCHER: *Drone sounds and environmental signals – a first review*. In P. BIRKHOLZ and S. STONE (eds.), *Proc. 30th ESSV Conference (Studientexte zur Sprachkommunikation, vol. 93)*, pp. 212–220. Dresden, Germany, March 2019.

[11] JOKISCH, O.: *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*, vol. 154, chapt. 25, pp. 305–317. Kursk, Russia, April 2019.

[12] JOKISCH, O., 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*, vol. 11658, pp. 180–190. Istanbul, Turkey, August 2019.

[13] voice INTER connect GmbH: *A development kit for the distant voice acquisition (vicDIVA)*. 2019. URL `www.voiceinterconnect.de/en/sdk_beamforming`. Retrieved 28/01/2020.

[14] ITU-T P.501: *Test signals for use in telephonometry, Recommendation P.501 (03/17)*. `https://www.itu.int/rec/T-REC-P.501-201703-I/en`, March 2017.

[15] JOKISCH, O., I. SIEGERT, and E. LOESCH: *Speech communication at the presence of unmanned aerial vehicles*. In *Proc. 46th Annual German Conference on Acoustics (DAGA 2020)*. Hannover, Germany, March 2020 (accepted).