PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Audiovisual Active Speaker Localization and Enhancement for Multirotor Micro Aerial Vehicles

Daniele Salvati, Carlo Drioli, Andrea Gulli, Gian Luca Foresti, Federico Fontana, Giovanni Ferrin

Department of Mathematics, Computer Science and Physics, University of Udine

{daniele.salvati, carlo.drioli, andrea.gulli, gianluca.foresti, federico.fonatana, giovanni.ferrin}@uniud.it

Abstract

We address the problem of localizing a speaker and enhancing his voice using audiovisual sensors installed on a multirotor micro aerial vehicle (MAV). Acoustic-only localization and signal enhancement through beamforming techniques is especially challenging in this conditions, due to the nature and intensity of disturbances originated by the electrical engines and the propellers. We propose a solution in which an efficient beamforming-based algorithm for both localization and enhancement of the source is paired to a video-based human face detection. The video processing front-end detects the human silhouettes and provides an estimation of direction of arrivals (DOAs) on the array. When the acoustic localization front-end detects a speech activity originating from one of the possible directions estimated by the visual components, the acoustic source localization is refined and the recorded signal is enhanced through acoustic beamforming. The proposed algorithm was tested on a MAV equipped with a compact uniform linear array (ULA) of four microphones. A set of scenes featuring two human subjects lying in the field of view and speaking one at a time is analyzed through this method. The experimental results conducted in stable hovering conditions are illustrated, and the localization and signal enhancing performances are analyzed.

Keywords: Audiovisual Speaker Localization, Speaker Enhancement, Diagonal Unloading Beamforming, Multirotor Micro Aerial Vehicle (MAV).

1 INTRODUCTION

The localization and signal enhancement of acoustic sources are important tasks in microphone array processing since many decades and they are of interest in applications such as teleconferencing, surveillance, animal ecology, human-computer interaction, and hearing aid [1, 2, 3, 4, 5]. They have also interesting application perspectives in a number of scenarios involving mobile robotic devices [6, 7, 8].

Attempts to tackle the acoustic related problems typical of multirotor aerial systems have been documented only recently [9, 10, 11, 12, 13, 14, 15, 16, 17]. When the acoustic recording is performed using microphone arrays installed on multirotor unmanned aerial vehicles (UAVs), the localization and signal enhancement of acoustic sources of interest becomes especially challenging, due to the number and variety of acoustic disturbances generated by this class of devices. Moreover, in the case of micro aerial vehicles (MAVs) of small size, the consequent constraints on the size of the microphone array may lead to poor signal-to-noise enhancement and poor spatial resolution issues. Beamforming and blind source separation are typical signal processing techniques used to this aim. In [13], the problem of enhancing a target speech source located in front of the UAV is solved using a minimum variance distortionless response (MVDR) beamformer coupled to a Wiener postfiltering scheme. The beamforming approach is the method of choice in a number of other investigations, e.g. [15]. When acoustic source localization is the only concern, localization methods with high-robustness with respect to low signal-to-noise ratios (SNRs) have also been investigated. In [17, 16] methods derived from the multiple signal classification (MUSIC) [18] were assessed, and the reported results show good localization performances. However, the method described also requires the monitoring of UAV inertial sensors and motor controls, and the learning or monitoring of propellers noise signal. Supervised noise suppression techniques relying on the online

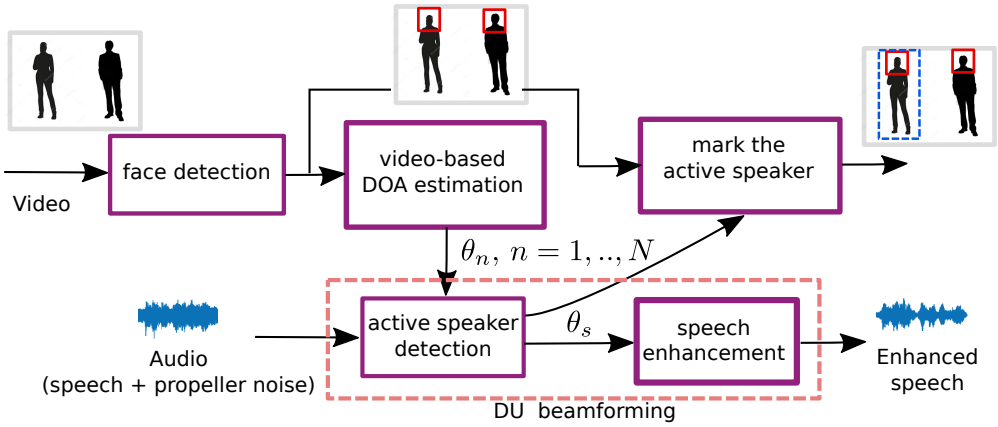

Figure 1. Schematic view of the processing workflow.

monitoring of the propellers noise, or on the off-line learning of its spectral and statistical characteristics, were also proposed in [19]. When, on the other hand, signal enhancement is the principal concern, time-frequency spatial filtering or blind source separation techniques have been used to selectively attenuate the propeller noise component while preserving the target acoustic source component. In [9], the time-frequency sparsity of specific target signals is exploited through the use of a time-frequency spatial filtering technique.

In this study, we address the problem of localizing a speaker and enhancing his voice using audiovisual sensors installed on a multirotor MAV. We propose a solution in which an efficient diagonal unloading (DU) beamforming-based algorithm [21] for both localization and enhancement of the source is paired to a video-based detection. The video processing front-end performs the human face detection and provides an estimation of direction of arrivals (DOAs) on the array. When the acoustic localization front-end detects a speech activity originating from one of the possible directions estimated by the visual components, the acoustic source localization is refined and the recorded signal is enhanced through acoustic beamforming. The audiovisual algorithm provides integrated source localization and signal enhancement capabilities, and provides superior spatial resolution properties when compared to conventional audio beamforming methods. The experimental results are conducted in stable hovering conditions and with different controlled signal-to-propeller-noise ratios (SPNRs).

2 DOA ESTIMATION AND SPEECH ENHANCEMENT

We address the problem of detecting the speech activity of a number of speakers positioned in front of a hovering drone, to decide who is the active speaker among them, estimate the DOA of the speech signal, and finally enhance the recorded speech through an acoustic beamforming directed toward the speech source.

We propose a solution in which an efficient beamforming-based algorithm for both the localization and enhancement of the source is paired to a video-based pattern recognition component. The processing scheme is illustrated in Figure 1. The video processing front-end performs human face detection and provides an estimation of DOAs of all possible speech sources. An array-based acoustic front-end is also active, which is capable of speech activity detection, localization and signal enhancement. When the acoustic localization front-end detects a speech activity originating from one of the possible directions estimated by the visual components, the acoustic source localization is refined and the recorded signal is enhanced through an acoustic beamformer.

The human face detection is based on a conventional Viola–Jones object detection algorithm [20], able to effectively detect one or more human face instances in real-time. Given the coordinates of the face bounding rectangles detected in the image frame and the angle of view of the video camera in use, an estimation of the DOA $\widehat{\theta}_n$, $n = 1, 2, \ldots, N$, (N is the number of speakers), corresponding to each possible source is computed.

The acoustic detection, localization and enhancement relies on the diagonal unloading (DU) beamforming-based algorithm [21, 22]. The DU beamformer is a data-dependent spatial filtering model that aims at exploiting the orthogonality property between signal and noise subspaces by subtracting an opportune diagonal matrix from the covariance matrix. The design and implementation of the DU beamformer is thus simple and effective, since it is obtained by computing the matrix (un)loading factor, providing high resolution directional response and noise robustness comparable to those of the MUSIC method and the MVDR beamformer while requiring less computational resources.

Consider a uniform linear array (ULA), the frequency-domain model of a typical acoustic narrowband beamformer, i.e. a spatial filter whose goal is to achieve directional signal reception, can be stated as

$$Y(k,f,\theta) = \mathbf{w}^H(k,f,\theta)\mathbf{x}(k,f), \tag{1}$$

where k is the block time index, f is the frequency bin, $\mathbf{x}(k,f)$ is the vector of signals captured by the array, $\mathbf{w}(k,f,\theta)$ is the beamformer coefficients for time-shifting, weighting, and summing the data so to steer the array in the direction θ, $Y(k,f,\theta)$ is the output of the narrowband beamformer, and H denotes the Hermitian (complex conjugate) transpose. The power spectral density of the spatially filtered signal is thus

$$P(k,f,\theta) = E\{|Y(k,f,\theta)|^2\} = \mathbf{w}^H(k,f,\theta)\mathbf{R}(k,f)\mathbf{w}(k,f,\theta), \tag{2}$$

where $\mathbf{R}(k,f) = E\{\mathbf{x}(k,f)\mathbf{x}^H(k,f)\}$ is the covariance matrix of the array signal ($E\{\cdot\}$ denotes mathematical expectation). The transformation on which the DU method is based, is obtained by subtracting an opportune diagonal matrix from the covariance matrix $\mathbf{R}(k,f)$ of the array output vector. The DU transformed matrix can be written as

$$\mathbf{R}_{\mathrm{DU}}(k,f) = \mathbf{R}(k,f) - \mu\mathbf{I}, \tag{3}$$

where μ is a real-valued, positive scalar, selected in such a way that the resulting covariance matrix is negative semidefinite in order to exploit the subspace orthogonality property. For more details, the reader can refer to [21, 22]. The DU procedure provides a solution for the signal enhancement beamforming coefficients $\mathbf{w}$:

$$\mathbf{w}_{\mathrm{DU}}(k,f,\theta) = \frac{\widehat{\mathbf{R}}_{\mathrm{DU}}(k,f)\mathbf{a}(f,\theta)}{\mathbf{a}^H(f,\theta)\widehat{\mathbf{R}}_{\mathrm{DU}}(k,f)\mathbf{a}(f,\theta)}, \tag{4}$$

where $\mathbf{a}(f,\theta)$ is the array steering vector and

$$\widehat{\mathbf{R}}_{\mathrm{DU}}(k,f) = \widehat{\mathbf{R}}(k,f) - \mathrm{tr}(\widehat{\mathbf{R}}(k,f))\mathbf{I}, \tag{5}$$

where $\mathrm{tr}(\cdot)$ is the operator that computes the trace of a matrix and $\widehat{\mathbf{R}}(k,f) = \mathbf{x}(k,f)\mathbf{x}^H(k,f)$ is the estimated covariance matrix that is computed with a single block due to the frame-by-frame computation of the signal enhancement beamforming. Given the beamforming coefficients, it is thus possible to compute the spatially filtered signal with respect to the estimated DOA, which will provide an enhanced version of the target source signal.

Let $\widehat{\theta}_n$ ($n = 1,2,\ldots,N$) denote the DOA of the n-th speaker estimated by the human face detection module. The voice activity detection (VAD) is calculated with an energy-based threshold:

$$\mathrm{VAD}(k) = \begin{cases} 1, & \text{if } \max\limits_{\widehat{\theta}_n}[\sum_{k_b=0}^{B-1}\sum_{f=f_{\min}}^{f_{\max}}|Y(k-k_b,f,\widehat{\theta}_n)|^2] > \eta, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

where $f_{\min}$ and $f_{\max}$ denote the frequency range for the beamforming computation, B is the number of frames considered for the speech detection, $Y(k-k_b,f,\widehat{\theta}_n) = \mathbf{w}_{\mathrm{DU}}^H(k-k_b,f,\widehat{\theta}_n)\mathbf{x}(k,f)$ is the signal enhancement DU beamforming for the direction $\widehat{\theta}_n$, and η is a given threshold. The parameter η was empirically set to the value allowing to effectively detect the source activity.

If the speaker detection returns an activity source $\mathrm{VAD}(k) = 1$, the steered response power (SRP) DU beamformer is computed to determinate the DOA of the active speaker. Specifically, we use a robust DU version [22] that is based on the estimation, computed by the power method, of the largest eigenvalue of the covariance matrix. The robust DU beamformer can be implemented by subtracting the largest eigenvalue from the diagonal elements of the estimated covariance matrix. This diagonal removal procedure allows the best suppression of signal subspaces. The robust DU beamforming for the localization is computed on the set of video-based DOAs $\widehat{\theta}_n$

$$P_{\mathrm{DU}}(k,f,\widehat{\theta}_n) = \frac{-1}{\mathbf{a}^H(f,\widehat{\theta}_\mathbf{n})(\widehat{\mathbf{R}}_l(k,f) - \widehat{\lambda}_1(k,f)\mathbf{I})\mathbf{a}(f,\widehat{\theta}_n)}, \tag{7}$$

where the covariance matrix for the localization $\widehat{\mathbf{R}}_l(k,f)$ is estimated through the averaging of the array signal blocks B

$$\widehat{\mathbf{R}}_l(k,f) = \frac{1}{B}\sum_{k_b=0}^{B-1}\mathbf{x}(k-k_b,f)\mathbf{x}^H(k-k_b,f), \tag{8}$$

and $\widehat{\lambda}_1(k,f)$ is the largest eigenvalue of $\widehat{\mathbf{R}}_l(k,f)$ that is estimated with the power method without adding significantly computational cost [22]. The $P_{\mathrm{DU}}(k,f,\widehat{\theta}_n)$ is related to the power contribution of a single frequency bin, and a function providing the broadband SRP-DU information of the whole frequency spectrum can be obtained by merging the contribution by some fusion criterion, such as the normalized frequency fusion proposed in [23]. The video-based broadband SRP-DU is thus defined as

$$P(k,\widehat{\theta}_n) = \sum_{\theta_\alpha=-\alpha}^{\alpha}\sum_{f=f_{\min}}^{f_{\max}}\frac{P_{\mathrm{DU}}(k,f,\widehat{\theta}_n+\theta_\alpha)}{||\mathbf{p}(k,f)||_\infty}, \tag{9}$$

where $||\cdot||_\infty$ denotes the uniform norm of the vector $\mathbf{p}(k,f) = [\sum_{\theta_\alpha=-\alpha}^{\alpha}P_{\mathrm{DU}}(k,f,\widehat{\theta}_1+\theta_\alpha),\sum_{\theta_\alpha=-\alpha}^{\alpha}P_{\mathrm{DU}}(k,f,\widehat{\theta}_2+\theta_\alpha),\ldots,\sum_{\theta_\alpha=-\alpha}^{\alpha}P_{\mathrm{DU}}(k,f,\widehat{\theta}_N)+\theta_\alpha)]$, and α is a DOA range parameter that takes into account the error resolution of the ULA. The DOA estimation of the speaker is provided by the maximum energy peak search

$$\hat{\theta}_s(k) = \underset{\widehat{\theta}_n}{\mathrm{argmax}}[P(k,\widehat{\theta}_n)]. \tag{10}$$

Finally, the enhanced speech signal is obtained by selecting the DU beamforming for the active speaker direction

$$Y(k-k_b,f,\hat{\theta}_s(k)) = \mathbf{w}_{\mathrm{DU}}^H(k-k_b,f,\hat{\theta}_s(k))\mathbf{x}(k-k_b,f), \quad k_b = 0,1,\ldots,B. \tag{11}$$

Note that if $\mathrm{VAD}(k) = 0$, then $Y(k-k_b,f,\hat{\theta}_s(k)) = 0$. By transforming $Y(k-k_b,f,\hat{\theta}_s(k))$ in the time-domain the enhanced speech signal is obtained.

To summarize, the overall procedure consists of:

- Video-based DOA estimation using the human face detection with the Viola–Jones algorithm [20];
- VAD estimation using the threshold-based signal power DU beamforming (6);
- Active speaker DOA estimation using the SRP-DU algorithm (10);
- Speech enhancement on the active DOA estimation using the DU beamforming (11).

3 EXPERIMENTAL SETUP AND RESULTS

The proposed algorithm was tested on a MAV equipped with a compact ULA of four microphones and a front video camera. The MAV used was a Parrot Bebop 1 quadcopter, with a 250 mm frame type, 400 g weight, and overall dimensions of 280 × 320 mm. The microphone array and video camera from a PlayStation Eye

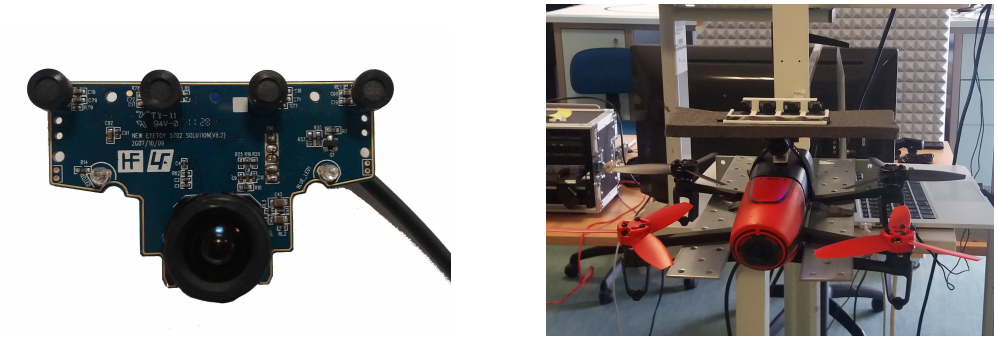

Figure 2. Left: the sensors board (four-microphone array and video camera); Right: the sensor board mounted on the quadcopter.

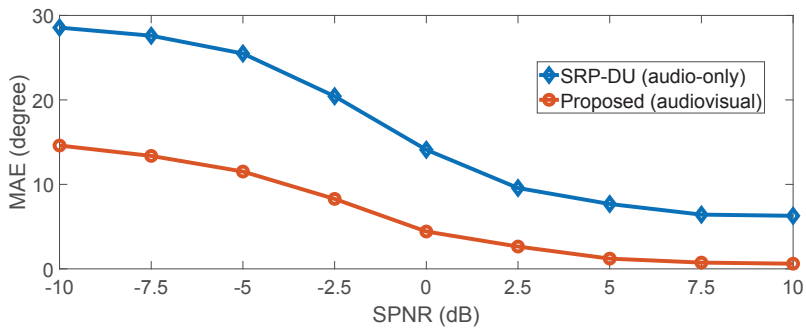

Figure 3. The active speaker localization performance at variation of SPNR.

USB device were used as the audiovisual front-end. This device provides a four-microphone uniformly spaced linear array with total size of 6 cm (the inter-microphone spacing being of 2 cm). The microphone array was fixed on the top of the MAV, centered with respect to the four propellers. A foam rectangular shield was put below the microphones to attenuate the direct path of the noise from the propellers. With this arrangement, the resulting DOAs of the four propellers are -30 degrees and 30 degrees. The angle of view of the video camera was approximately of 60 degrees. The audiovisual recording was performed by a dedicated computer positioned next to the MAV. A picture of the MAV setup is provided in Figure 2. The video resolution was 640 $\times$ 480, at a frame rate of 50 fps. The audio sampling frequency was 16 kHz, and the block size was 2048 samples with an overlap of 512 samples. A Hann window was used. The spatial resolution was set to 1 degree in the range [-30, 30] degrees which corresponds to the field of view of the video camera. The number of frames for the speaker detection and localization was $B = 25$. The parameter η (6) for the VAD was 10^{-8}. The parameter α (9) for the acoustic localization was set to 3. The frequency range was [100, 6000] Hz.

A set of audiovisual recordings of two subjects, standing in the field of view and speaking one at a time at different positions with respect to the MAV, was collected. The recordings were performed with the MAV in hovering conditions, thus the speech was corrupted by the propeller noise in realistic acoustic conditions. The two speakers were positioned symmetrically with respect to the frontal axis of the camera, uttering the same sentence one at the time. The subjects were instructed to speak one at a time and to hold their position until both uttered the same predetermined sentence. Then, they were asked to move to the next position. The recording positions where 15 in total, according to the following distances and angles: 2, 3, and 4 m at [-5, 5], [-10, 10], [-15, 15], [-20, 20], and [-25, 25] degrees.

Figure 3 shows the localization performance at variation of SPNR expressed as mean absolute error (MAE) of the proposed audiovisual system compared to an audio-only SRP-DU beamformer [22]. We can observe

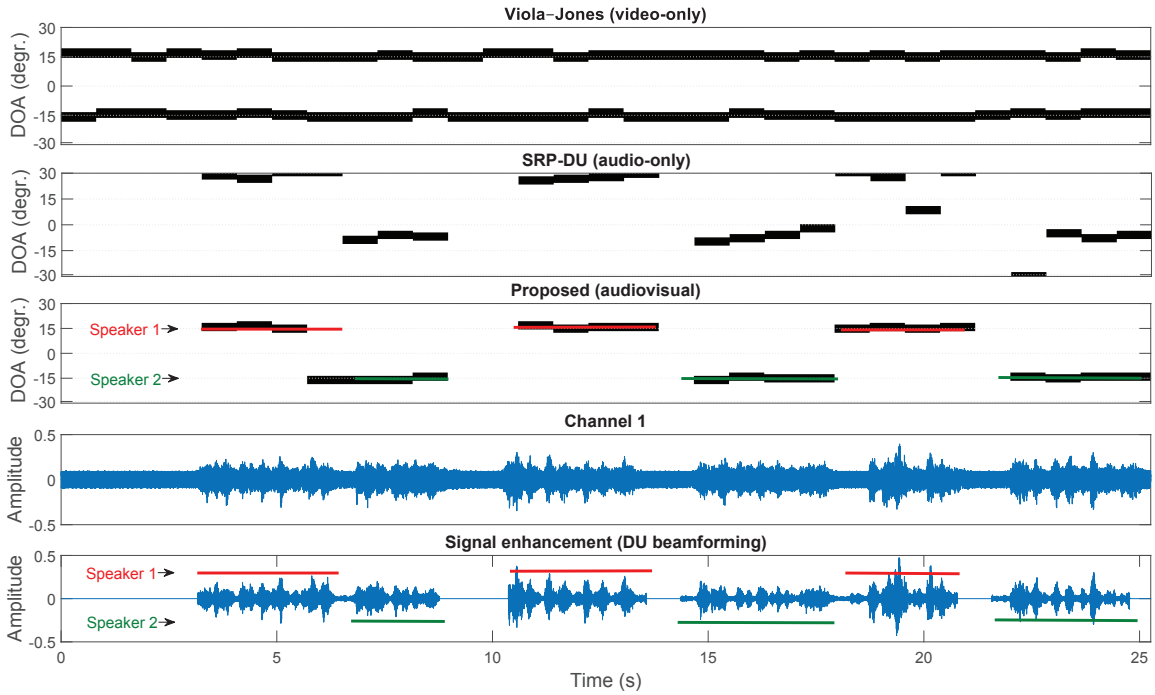

Figure 4. An example of detection, localization, and enhancement performance of two subjects, speaking one at a time at different positions (2 m from the array at [-15, 15] degrees) with an SPNR of 0 dB.

Table 1. The output SPNR of the signal enhancement with the conventional DS and the DU beamformer.

SPNR (dB)	10	5	0	-5	-10
DS	15.04	10.05	4.97	-0.36	-6.65
DU	28.55	23.63	18.79	14.32	10.70

the better MAE of the proposed system, and the poor accuracy of the audio-only system. Figure 4 depicts an example of detection, localization, and enhancement performance for the case of two subjects at 2 m from the array and at [-15, 15] degrees with an SPNR of 0 dB. The video-only DOA estimation has an average MAE of 1 degree. We can note the correct detection and localization of the proposed system, expected for the case of the last part of the first sentence of speaker 1 in which the DOA is not correctly associated to the true speaker. We can also see the signal enhancement with the attenuation of the drone propeller noise. The speaker detection activity performance is reported in Figure 5 in terms of percentage of detection rate (DR). The DR decreases at low SPNR, and it is accurate up to an SPNR of 0 dB. Finally, Table 1 shows the speaker enhancement performance of the DU beamformer compared to the conventional delay-and-sum (DS) beamformer [24]. The performance is reported comparing the output SPNR of the signal enhancement at different input SPNR conditions. The DU beamformer significantly outperforms the conventional one.

4 CONCLUSIONS

We have presented a system able of localizing a speaker and enhancing his voice using audiovisual sensors installed on a multirotor MAV. We have proposed a solution in which an efficient DU beamforming-based algorithm for detection, localization and enhancement of a speaker is paired to a video-based detection. The video

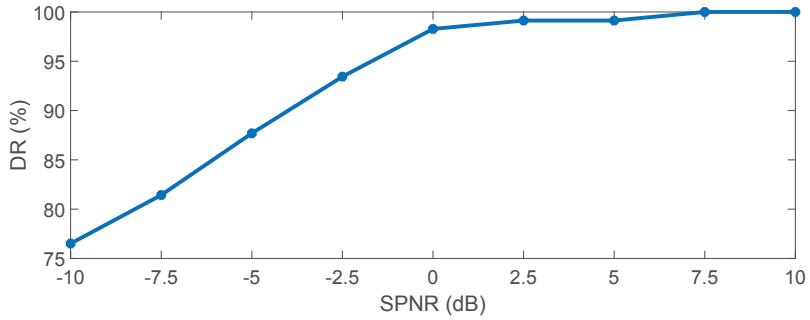

Figure 5. The VAD performance of the proposed audiovisual system at variation of SPNR.

processing front-end performs a human face detection and provides an estimation of DOAs that are used to refine the speaker activity acoustic detection and localization, and the final speech enhancement. The proposed algorithm was tested on a MAV equipped with a compact ULA of four microphones. The experimental results conducted in stable hovering conditions have shown that the proposed audiovisual system significantly improves the DOA estimation if compared to an audio-only approach. We have shown that the signal enhancement DU beamforming provides a better output SPNR if compared to the conventional DS beamformer.

REFERENCES

[1] D. Salvati, C. Drioli, and G. L. Foresti, "A weighted MVDR beamformer based on SVM learning for sound source localization," *Pattern Recognition Letters*, vol. 84, pp. 15–21, 2016.

[2] J. Traa, D. Wingate, N. D. Stein, and P. Smaragdis, "Robust source localization and enhancement with a probabilistic steered response power model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 493–503, 2016.

[3] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wireless Communications and Mobile Computing*, vol. 2017, pp. 1–24, 2017.

[4] D. Salvati, C. Drioli, and G. L. Foresti, "Exploiting a geometrically sampled grid in the steered response power algorithm for localization improvement," *Journal of the Acoustical Society of America*, vol. 141, no. 1, pp. 586–601, 2017.

[5] C. Drioli, G. Giordano, D. Salvati, F. Blanchini, and G. L. Foresti, "Acoustic target tracking through a cluster of mobile agents," *IEEE Transactions on Cybernetics*, pp. 1–14, 2019.

[6] J. M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, vol. 2, pp. 1228–1233.

[7] R. Levorato and E. Pagello, "DOA acoustic source localization in mobile robot sensor networks," in *Proceeding of the IEEE International Conference on Autonomous Robot Systems and Competitions*, 2015, pp. 71–76.

[8] M. Basiri, F. Schill, P. Lima, and D. Floreano, "On-board relative bearing estimation for teams of drones using sound," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 820–827, 2016.

[9] L. Wang and A. Cavallaro, "Microphone-array ego-noise reduction algorithms for auditory micro aerial vehicles," *IEEE Sensors Journal*, vol. 17, no. 8, pp. 2447–2455, 2017.

[10] P. Nguyen, H. Truong, M. Ravindranathan, A. Nguyen, R. Han, and T. Vu, "Matthan: Drone presence detection by identifying physical signatures in the drone's RF communication," in *Proceeding of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 211–224.

[11] K. Hoshiba, K. Washizaki, M. Wakabayashi, T. Ishiki, M. Kumon, Y. Bando, D. Gabriel, K. Nakadai, and H. G. Okuno, "Design of UAV-embedded microphone array system for sound source localization in outdoor environments," *Sensors*, vol. 17, no. 11, 2017.

[12] S. Oh, Y.-J. Go, J. Lee, and J.-S. Choi, "Sound source positioning using microphone array installed on a flying drone," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3422–3422, 2016.

[13] Y. Hioka, M. Kingan, G. Schmid, and K. A. Stol, "Speech enhancement using a microphone array mounted on an unmanned aerial vehicle," in *Proceeding of the IEEE International Workshop on Acoustic Signal Enhancement*, 2016, pp. 1–5.

[14] S. Argentieri, P. Danes, and P. Soueres, "A survey on sound source localization in robotics: from binaural to array processing methods," *Computer Speech and Language*, vol. 34, no. 1, pp. 87–112, 2015.

[15] T. Ishiki and M. Kumon, "Design model of microphone arrays for multirotor helicopters," in *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 6143–6148.

[16] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, K. Nakadai, and H. G. Okuno, "Noise correlation matrix estimation for improving sound source localization by multirotor UAV," in *Proceeding of the International Conference on Intelligent Robots and Systems*, 2013, pp. 3943–3948.

[17] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, "Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter," in *Proceeding of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 3288–3293.

[18] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[19] S. Yoon, S. Park, Y. Eom, and S. Yoo, "Advanced sound capturing method with adaptive noise reduction system for broadcasting multicopters," in *Proceeding of the IEEE International Conference on Consumer Electronics*, 2015, pp. 26–29.

[20] P. Viola and M. J."Jones, Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 2004.

[21] D. Salvati, C. Drioli, and G. L. Foresti, "A low-complexity robust beamforming using diagonal unloading for acoustic source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 609–622, 2018.

[22] D. Salvati, C. Drioli, and G. L. Foresti, "Power method for robust diagonal unloading localization beamforming," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 725–729, 2019.

[23] D. Salvati, C. Drioli, and G. L. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 581–585, 2014.

[24] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.