BEAMFORMING WITH KINECT V2

Stefan Gombots*, Felix Egner†, Manfred Kaltenbacher‡
Institute of Mechanics and Mechatronics, Vienna University of Technology
Getreidemarkt 9, 1060 Wien, AUT
*e–mail: stefan.gombots@tuwien.ac.at
†e–mail: felix.egner@tuwien.ac.at
‡e–mail: manfred.kaltenbacher@tuwien.ac.at

Abstract – Microphone array measurements in combination with beamforming techniques are often used for acoustic source localization. The sound pressure obtained at different microphone positions are mapped by these techniques to a planar or a surface map. The mapping result named as beamform map, indicates the location and strength of acoustic sources. For this mapping process the distance between the sound source or device under test (DUT), and the microphone positions must be known. To determine these distances the Microsoft Kinect for Windows v2 (Kinect V2) is used. The Kinect V2 sensor allows acquiring RGB, infrared (IR) and depth images. The depth images are evaluated and the required distances are computed. The distance is measured contactless, and also the surface of the DUT can be reconstructed through the depth images. Furthermore, the RGB image is used as an underlying layer of the beamform map. The applicability of the source mapping process using the Kinect V2 is demonstrated and the characteristics of the sensor are discussed.

Keywords – Beamforming, microphone array, source mapping, Kinect V2

I. INTRODUCTION

Beamforming techniques, e. g. Standard Beamforming [1], Functional Beamforming [2], CLEAN SC [3], Orthogonal Beamforming [4], are an often used method to localize acoustic sources. These techniques are based on evaluating simultaneously collected sound pressure data from microphone array measurements. In the case of stationary acoustic sources it is common to work in the frequency domain. Here, the beamform map for Standard Beamforming is computed by

$$a(\boldsymbol{g}) = \boldsymbol{g}^H \boldsymbol{C} \boldsymbol{g} \tag{1}$$

with $\boldsymbol{g}$ the steering vector, ^{H} the hermitian operation, and $\boldsymbol{C}$ the cross spectral matrix of the microphone signals. The beamform map provides information about location and strength of sound sources. Thereby, a certain model for the sources and sound field is assumed. By using monopole sources in a free field, the steering vectors $\boldsymbol{g}$ are given by the free-space Green's function

$$g(r) = \frac{1}{4\pi r} e^{-jkr} \tag{2}$$

with the wave number k and $r = |\boldsymbol{x}_\mathrm{s} - \boldsymbol{x}_{\mathrm{m},i}|$ the distance between assumed source point $\boldsymbol{x}_\mathrm{s}$ and microphone position $\boldsymbol{x}_{\mathrm{m},i}$. The typical geometric setup of acquiring a two-dimensional beamform map is depicted in Fig. 1. The sound sources are assumed to be in a planar scanning area. The microphones are located parallel to this area. Hence, in the two-dimensional planar source mapping the z-coordinate is constant. Now,

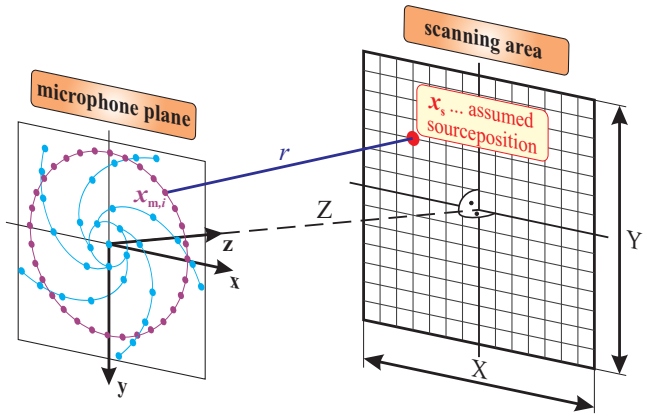

Fig. 1: GEOMETRIC SETUP – Two-dimensional acoustic source mapping.

using the depth image information of the Kinect V2 sensor, the source distribution can be mapped on the real surface of the DUT. In addition, the information of the RGB image can be used as an underlying layer of the beamform map.

II. MEASUREMENT SYSTEM

In the following the measurement system for obtaining the sound pressure at the microphone positions $\boldsymbol{x}_{\mathrm{m},i}$ will be presented and the characteristics of the Kinect V2 sensor discussed. In Fig. 2, one can see a schematic representation of the overall measurement system used, containing the microphone array, the data acquisition unit and the Kinect V2.

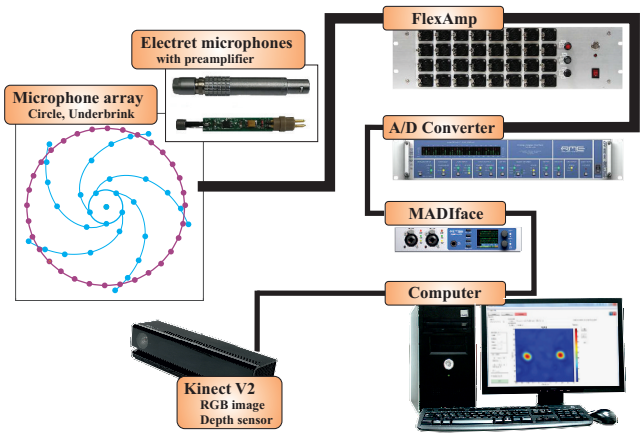

Fig. 2: Measurement system – Schematic representation.

A. *Microphone array*

The result of the acoustic source mapping process depends on different parameters, e. g. the microphone arrangement, the frequency of the acoustic source, the signal processing algorithm, etc. The used planar array has 63 electret microphones and consists of a circle with 32 microphones and an Underbrink design [5] of 31 microphones. The aperture of the array is 1 m, resulting in a lower frequency limit of about 1400 Hz. According to the spatial aliasing theorem, which is introduced by the repeated sampling space of the microphones, the upper limit of the circle array is about 8000 Hz. At higher frequencies ghost images will arise in the beamform map. The theorem holds for regular arrays, like the circle. Irregular array designs, where the microphone spacings are different, the effect of ghost images can be decreased. In this context the Underbrink design performs best [6]. Within the scope of this work, all 63 microphones are used to calculate the beamform map.

The electret microphones are calibrated in an anechoic chamber by comparing it with a calibrated Bruel&Kjaer microphone. The calibration process was also verified using a pistonphone. The sensitivities of the 63 microphones are considered in the calculation process of the beamform map. In beamforming methods the microphone array is usually placed in the far field of the source. Moreover, the source should be kept near the center axis of the array for best results. Hence, the directional characteristic of the microphones seems to be negligible and therefore are not considered.

The control of the recording, the analysis, the signal processing and the computation of the beamform map is done by a program written in MATLAB.

B. *Kinect V2 sensor*

With the Kinect V2 sensor, Microsoft delivers an interface device for their gaming console Xbox One, providing a motion controller and a speech recognition system. With an adapter [7] the sensor can also be used by computer to acquire RGB and depth images. To enable the use of the Kinect V2 one has to download the Kinect for Windows SDK 2.0 (free available). It provides the drivers, application programming interfaces (APIs) and code samples. Since Version 2016a the Kinect V2 is also supported by MATLAB. To use the functionality of the sensor in earlier versions one can use the Kin2 Toolbox [13]. Both tools uses the underlying functions of the SDK.

The Kinect V2 is composed of a RGB and an IR camera, whereby the IR camera is used for the acquisition of the depth images. Figure 3 illustrates the component parts of the sensor.

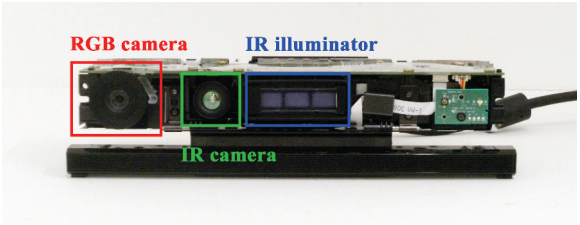

Fig. 3: Kinect V2 sensor – Component parts.

The sensor uses the time-of-flight method (ToF) for the depth measurement; a detailed description is given in [8]. Some characteristics of the sensor are summarized in Tab. 1.

RGB camera	resolution	1920 × 1080 px
	FOV (h × v)	84.1° × 53.8°
IR/depth camera	resolution	512 × 424, px
	FOV (h × v)	70.6° × 60.0°
	operating range	0.5 – 4.5 (8) m
Frame rate		30 Hz
Connection type		USB 3.0

Tab. 1: Characteristics of the Kinect V2 sensor [9].

The field of view (FOV) of both cameras have been checked. For this reason the sensor was placed parallel to a white plane wall in a distance of 1 m and 2 m. The result is given in Tab. 2. The measurement shows a good agreement with the specification given by the manufacturer.

		Manufacturer	Own evaluation
RGB camera	h	84.1°	85.0°
	v	53.8°	54.2°
IR/depth camera	h	70.6°	70.3°
	v	60.0°	58.2°

Tab. 2: Field of view – Comparison.

Because the RGB and depth image have different field of view, the correspondence between the images have to be established. In Fig. 4 the difference in the FOV of

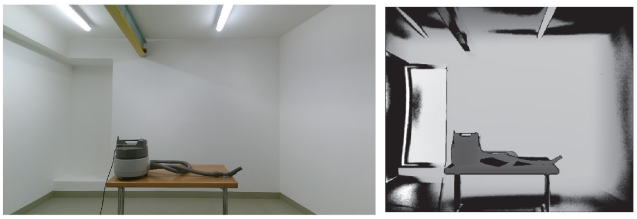

Fig. 4: Images from Kinect V2 – (left) RGB image (right) depth image.

the cameras is shown. One can see that the images partially overlap, but the color camera has a wider horizontal FOV, while the IR camera has a larger vertical FOV. The correspondence between the images can be established by the SDK functions. Making use of the SDK functions, the mapping between the locations on the depth image and their corresponding locations on the color image can be done. The Kinect V2 sensor is placed near the center of the array to exploit the FOV of the cameras best (Fig. 5). The base of the Kinect V2 was removed to place it on the given array geometry.

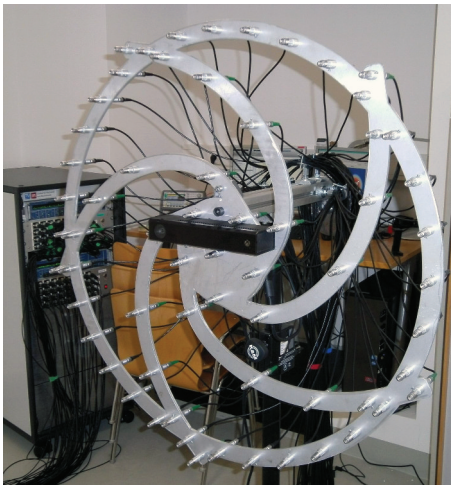

Fig. 5: Array geometry with Kinect V2.

Next, some influences on the depth images are discussed. Previous investigations have shown that the Kinect V2 need a pre-heating time before providing accurate range measurements [10]. After 20 minutes of usage the distance variation becomes nearly constant (more or less 1 mm).

The depth images oscillate during the measurement, known as wiggling error of ToF cameras [11]. Therefore the depth images have been averaged to decrease this effect. The decrease of the depth fluctuations by averaging is shown in Fig. 6. One can state, that an averaging of at least 50 frames should be done to get accurate results.

Next the deviation between measured depth and the real distance was determined. For this, the sensor was placed parallel to a plane white wall at several distances.

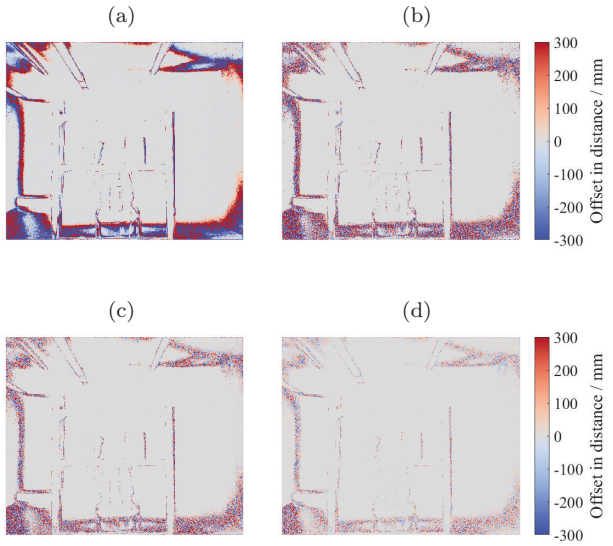

Fig. 6: Improvement of the depth map through averaging – Result of (a) no (b) 10 (c) 20 (d) 50 averages. Offset in distance to an averaged map of 100 frames.

Depth measurements were taken and averaged over 50 frames. To determine the mean value a small section (50 px × 50 px) at the image center was used. The real distances were measured by a laser distance meter and also by a tape measure with an accuracy of about ±2 mm. The deviation between the mean value and the true distances is given in Fig. 7. Depth measurements among 800 mm seems not to be suitable.

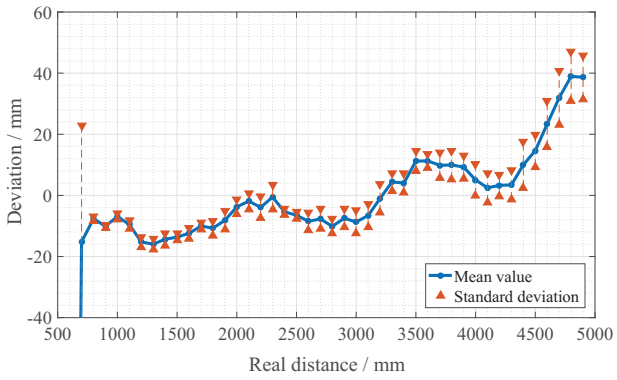

Fig. 7: Depth variation of measured and true distances.

Furthermore, using the same setup the standard deviation for each pixel of the depth images were computed. The measurements show that with increasing distance the errors towards the edges increase (see Fig. 8). Therefore, to make accurate long-distance measurements, the DUT should be placed in the center of the image.

Further influences are given by the albedo of surfaces.

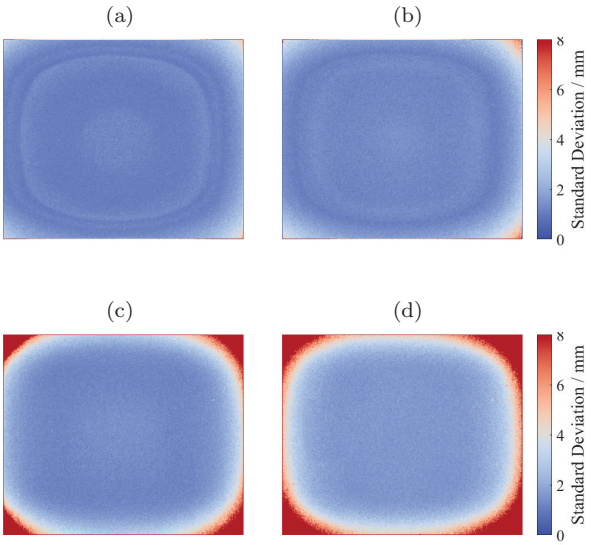

Fig. 8: Standard deviation of each pixel – Considerd distances (a) 800 mm (b) 1000 mm (c) 1500 mm (d) 2000 mm, averaging 50 frames.

It has been shown that on very reflective as well as very dark surfaces the corresponding distances in the depth images are larger then expected [10]. To overcome this limitation the supposed surfaces in the experimental setup (reflective and dark) were covered by a developer spray. Especially the reflective surface of the hairdryer in the experimental setup leads to depth errors (see Fig. 9).

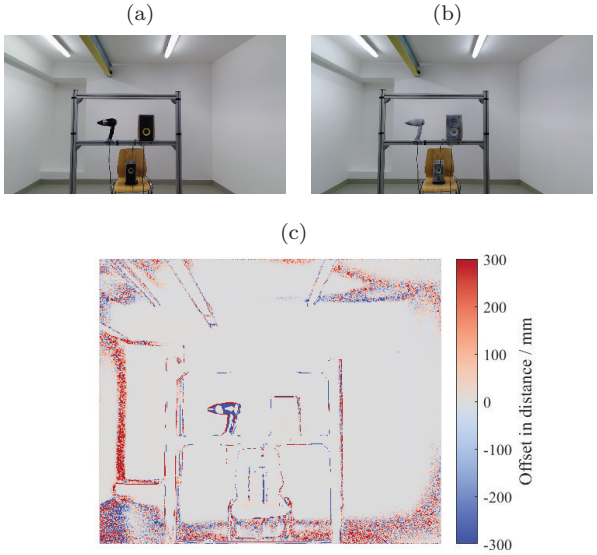

Fig. 9: Albedo influence – (a) Experimental setup (b) setup after covering the surfaces by developer spray (c) depth errors between (a) and (b) of an averaged snapshot of 50 frames.

III. CALIBRATION

For underlying the beamform map by the RGB image the camera was calibrated using the algorithm described in [12]. To this purpose a checkerboard is placed in different positions in front of the camera. The dimensions of the checkerboard pattern have to be known. For best results 10 to 20 images of the pattern should be used. The images should be taken at a distance approximately equal to the distance from the camera to the DUT. As a rule of thumb, the checkerboard should cover at least 20 % of the image. As one can see in Fig. 10 the checkerboard should also be captured at different orientations.

Fig. 10: RGB calibration – Different positions and orientations of the checkerboard.

At the end of the calibration process, one obtains the intrinsic and extrinsic camera parameters. The intrinsic parameters of the RGB camera are compared to a self calibration implemented in the Kin2 toolbox [13]. For the IR camera, the SDK provides a function which delivers the intrinsic parameters. The same calibration method, which was used for the RGB camera, can be also applied to the IR camera. In Tab. 3 the intrinsic parameters of the RGB and the IR are listed and compared.

	RGB camera		IR camera	
	[13]	self	SDK	self
f_x (px)	1063.86	1110.25 $\pm$ 95.84	366.8731	365.62 $\pm$ 17.84
f_y (px)	1063.86	1135.20 $\pm$ 98.49	366.8731	373.95 $\pm$ 17.77
c_x	978.54	953.58 $\pm$ 15.16	259.78	254.09 $\pm$ 2.85
c_y	535.62	539.22 $\pm$ 17.83	208.02	254.09 $\pm$ 4.98
K_1	0.01849	0.05550 $\pm$ 0.03501	0.09639	0.12603 $\pm$ 0.04531
K_2	-0.01016	-0.01286 $\pm$ 0.08980	-0.27167	-0.34688 $\pm$ 0.16376
K_3	0.01006	-0.04601 $\pm$ 0.06957	0.08992	0.16029 $\pm$ 0.15585

Tab. 3: RGB und IR camera intrinsics – Focal length (f_x, f_y), principal point (c_x, c_y), radial distortion coefficients (K_1, K_2, K_3).

Knowing these parameters the lens distortion can be corrected, see Fig. 11. The parameters are quite sensitive to the selected images for the calibration. However, for the purpose of this work the factory-set calibration values are used. As previously mentioned the SDK pro-

vides also a coordinate mapping between RGB and IR image which should be used.

Fig. 11: Lens distortion – (left) distorted RGB image (right) undistorted RGB image.

The depth images were corrected by using the information of Fig. 7. For overlaying the RGB image with the beamform map some corrections have to be done. Due to the fact, that the y-axis of the Kinect V2 doesn't coincide with the microphone plane, a 1.8 degree rotation about the x-axis was done. Reason for that could be a non-perfect attachment of the sensor. Furthermore, the origin of the Kinect V2 coordinate system doesn't match the origin of the array coordinate system, since the Kinect V2 isn't placed in the center of the array. The beamform map has to be shifted in the y and -x direction. The translation in y direction is 80 mm and in the -x direction 120 mm.

IV. EXPERIMENTAL RESULTS

Experiments with real sources are made to demonstrate the acoustic source mapping using the RGB and IR images. For this purpose two smallband and a broadband noise source were used (two speakers and a hairdryer, see Fig. 12). Speaker 1 and the hairdryer are almost at the same distance to the microphone plane. Speaker 2 is approximately 50 cm behind them. The

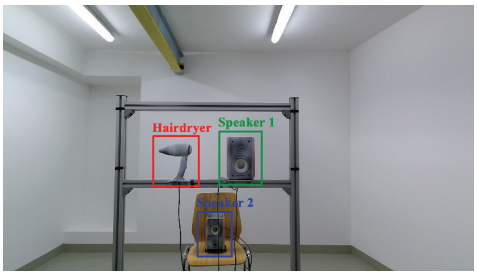

Fig. 12: Experimental setup.

sampling frequency has been 48 kHz, the measurement time 5 s and the temperature 25° C. The frequency spectrum of the center microphone and the noise level is depicted in Fig. 13. The spectrum was averaged 100 times using the Hanning window and a block size of 4096 samples with a block overlapping of 50 %.

To identify sound sources one choose the characteristic peaks of the spectrum. For obtaining the beamform

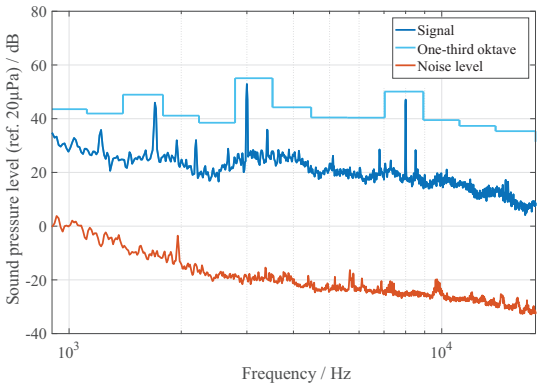

Fig. 13: Frequency spectrum of the center microphone.

maps a one-third octave analysis was done. First the two-dimensional mapping using a constant distance Z between microphone plane and scanning area was chosen. Then the beamform maps were put on the RGB image of the Kinect V2. The results are given in Fig. 14. The beamform maps were normalized to the maximum.

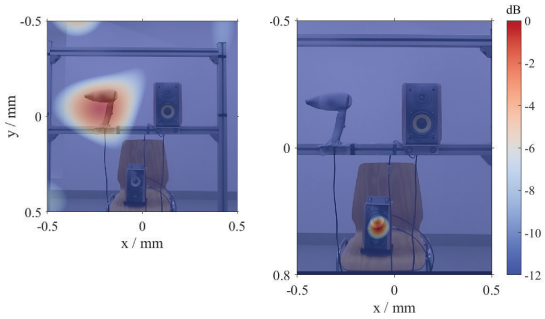

Fig. 14: Beamform maps – one-third octave analyses (left) 1600 Hz, $Z = 1400$ mm (right) 8000 Hz, $Z = 1900$ mm.

Next the depth informations of the sensor should be used. There are two possible ways to use them. First, the depth image can be used as a weighting of the two-dimensional beamform map. To do that, the beamform map will be calculated in a normal way using a constant Z. Then this result will be mapped on the 3 dimensional scene of the depth image. Second, the depth information can directly be used in Eq. 2 as assumed source point $\boldsymbol{x}_s$, meaning that the steer vectors depends on the Kinect V2 measurements. The mapping process of both ways is shown in Fig. 15. In the surface mapping the depth informations inside 1.3 m and 1.5 m were used. To show the difference between both methods the surface map was projected on a plane (see Fig. 16). Both methods provide the sound source in the one-third octave of 1600 Hz (hairdryer). Differences in the beamform maps are given through the different distances used for the calculation of the steer vectors.

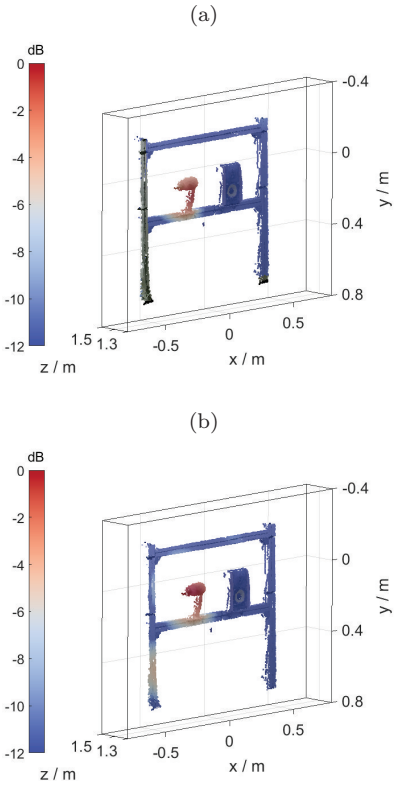

Fig. 15: One-third octave (1600 Hz) – Surface map of (a) method 1 and (b) method 2.

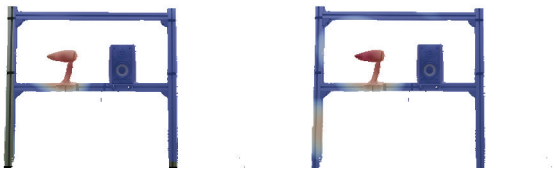

Fig. 16: Comparision – (left) method 1 (right) method 2.

V. CONCLUSION

The applicability to use the Kinect V2 sensor on beamforming was demonstrated. The surface mapping process provided good results. The different effects on the acquired depth images were examined. These yields to some corrections on the depth image. Moreover the overall measurement system was presented. Further investigations should be done to see, if a manual calibration and the point cloud acquisition (mapping between the depth and RGB image) can enhance the accuracy.

VI. REFERENCES

[1] Th. Mueller, "Aeroacoustic Measurements", *Springer*, ISBN 3-540-41757-5, 2002.

[2] R. Dougherty, "Functional Beamforming", *5th Berlin Beamforming Conference*, BeBeC-2014-01, 2014.

[3] P. Sijtsma, "Clean based on spatial source coherence", *Int. J. Aeroacoustics 6*, pp 357-374, 2009.

[4] E. Sarradj, "A fast signal subspace approach for the determination of absolute levels from phased microphone array measurements", *Journal of Sound and Vibration 329*, pp 1553-1569, 2010.

[5] J.R. Underbrink, *Circularly symmetric, zero redundancy, planar array having broad frequency range applications*, Pat. US6205224 B1, 2001.

[6] Z. Prime and C. Doolan, "A comparision of popular beamforming arrays", *Proceedings of ACOUSTICS*, 2013

[7] `https://www.microsoftstore.com/store/msusa/en_US/pdp/Kinect-Adapter-for-Xbox-One-S-and-Windows-PC/productID.2233937600`.

[8] J. Sell and P. O'Connor, "The Xbox One System on a Chip and Kinect Sensor", *IEEE Micro*, vol 34, no. 2, pp 44-53, 2014.

[9] P. Fankhauser, M. Bloesch, D. Rodriguez, R. Kaestner, M. Hutter, R. Siegwart, "Kinect v2 for Mobile Robot Navigation: Evaluation and Modeling", *2015 International Conference on Advanced Robotics (ICAR)*, IEEE, pp 388-394, 2015.

[10] E. Lachat, H. Macher, M.-A. Mittet, T. Landes, P. Grussenmeyer, "First experiences with Kinect V" sensor for close range 3D modelling *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol XL-5/W4, pp 93-100, 2015

[11] S. Foix, G. Alenya and C. Torras, "Lock-in Time-of-Flight (ToF) Cameras: A Survey", *IEEE Sensors Journal*, vol 11, no. 3, pp 1-11, 2011.

[12] Z. Zhang. "A flexible new technique for camera calibration", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330-1334, 2000.

[13] J. R. Terven and D. M. Córdova-Esparza, "Kin2. A Kinect 2 toolbox for MATLAB", *Sience of Computer Programming*, vol 130, pp 97-106, 2016, `http://dx.doi.org/10.1016/j.scico.2016.05.009`