# Head Orientation Estimation from Multiple Microphone Arrays

Rebecca C. Felsheim*†[1] , Andreas Brendel*, Patrick A. Naylor†, Walter Kellermann*
* *Multimedia Communications and Signal Processing Lab, Friedrich Alexander Universität Erlangen-Nürnberg*
Email: {rebecca.felsheim, andreas.brendel, walter.kellermann}@fau.de
† *Electrical and Electronic Engineering, Imperial College London*, Email: p.naylor@imperal.ac.uk

***Abstract*—Knowledge of head orientation is important for various audio signal processing tasks involving human speakers, including speech enhancement and attention tracking. Most of the methods estimate the head orientation using video information which, however, is not always available. In this work, two known audio features for head orientation estimation are reviewed and three new features are proposed. Furthermore, all evaluated features have been combined in two different ways: with a linear combination and a small artificial neural network. The resulting algorithms are able to detect the head orientation in our experiments with high precision and show superior performance over state-of-the-art methods.**

***Index Terms*—acoustic head orientation estimation, OGCF, HLBR, CDR**

## I. Introduction

An estimate of head orientation for a human speaker as a directional sound source is useful for various signal processing tasks, e.g., to track the attention of a person [1]. As the sound is radiated and distorted differently depending on the orientation of the head relative to the observing microphones, knowledge of the head orientation of a speaker can be used to enhance recorded speech accordingly [2] or to support speaker localization algorithms [3].

The head orientation can either be estimated from acoustic or visual data, where the latter is a much more mature field of research. As an example, the survey [4] shows a large variety of approaches using visual input for the estimation of some or all of the three degrees of freedom of head orientation.

If, additionally, audio data is available, the performance can be improved compared to the use of only visual input [1]. In [5], video and audio data have been fused to estimate the speaker's position in addition to its orientation, and both position and head orientation are tracked.

However, very often visual data is not available and, for some tasks such as speech enhancement and source localization, it is not always necessary. Sometimes, even if it were a technical option, it is not usable, e.g., because of privacy issues. In any case, the joint use of acoustic and visual data requires additional calibrated camera hardware synchronized with the microphone hardware. Additionally, the performance of algorithms based on video data is degraded by insufficient lighting or objects obstructing the view [6]. Therefore, reliable head orientation estimation using only acoustic data is desirable in many situations.

Some of the early work on acoustic head orientation estimation [7] estimated the head orientation using the energy of the lowpass-filtered signals from 448 microphones distributed along the walls of a shoebox-like enclosure. The Oriented Global Coherence Field (OGCF) method proposed in [8], which is based on the Generalized Cross-Correlation Phase Transform (GCC-PHAT) [9] estimates the head orientation using five T-shaped microphone arrays with 4 microphones each, distributed along three walls of a room. The GCC-PHAT-related Steered Response Power PHAT (SRP-PHAT) has been used by [3] to estimate the head orientation. That work compared the SRP-PHAT-based approach with the High-to-Low-Band Energy Ratio (HLBR) head orientation algorithm and showed that the SRP-PHAT-based approach is slightly better than the HLBR method. In [6], with a single 16 element linear microphone array, the use of the HLBR, the received signal energy and the Direct-to-Diffuse Speech Ratio (DDR) were proposed for the head orientation estimation. Recent work on acoustic head orientation estimation [10] uses the power of Mel bands normalized to the average power over all microphones. For the classification of the orientations a neural network is compared to a simple mapping table. The accuracy of both algorithms is similar but the mapping table classifies less orientations than the neural network.

In this work, we focus on the estimation of the head orientation in the azimuthal plane. For this, three new features are proposed. One of them is based on the Coherent-to-Diffuse Power Ratio (CDR) [11]. The other two, High Band Variance (HBV) and the Spectral Difference (SD), exploit the shaping of the high frequency spectrum by the head radiation. The new features are then compared against two existing features OGCF [8] and HLBR [3]. Furthermore, two combinations of all five features are evaluated. The first one combines the features linearly, while the second uses a small artificial neural network which is trained for a specific room. As the different features extract different characteristics of the microphone signals, which complement each other for head orientation estimation, their combinations increase the robustness against noise and fewer microphones are necessary for similar accuracy.

The remainder of the paper is structured as follows. The

This work was supported by DFG under contract no <Ke890/10-1> within the Research Unit FOR2457 "Acoustic Sensor Networks".

[1]Rebecca Felsheim was a visiting student at the Imperial College London while conducting this research.

known features for head orientation estimation are reviewed in Section II and the proposed features are presented in Section III. The head orientation estimation methods are introduced in Section IV. In Section V the experimental evaluation and the used speech data are described and in Section VI the results are discussed.

## II. State-of-the-art Acoustic Features

Human talkers do not radiate signal energy uniformly in all directions but more to the front than to the sides, the top or the back of the talker [12]. Furthermore, the directionality of the sound radiation increases with frequency. For low frequencies up to 500 Hz the energy radiation is nearly uniform for all directions, while for higher frequencies the radiation pattern tends towards a cardioid shape [8]. However, the exact radiation pattern of human speakers cannot be described in closed form due to the variations between the human subjects and the acoustic conditions.

In the following all algorithms are described for $M$ microphone arrays with $Q$ microphones each, distributed on the walls of a room. It is assumed that the location $\boldsymbol{s} = (x_s, y_s, z_s)$ of the speaker's mouth is known.

### A. Oriented Global Coherence Field

The OGCF algorithm [8], [13] is based on the GCC-PHAT [9], which is defined as

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{X_1(k) X_2^*(k)}{|X_1(k) X_2^*(k)|} e^{j\omega_k \tau} d\omega_k \ , \qquad (1)$$

where $\tau$ describes the time lag and $\omega_k$ the angular frequency corresponding to frequency bin $k$. $X_1(k)$ and $X_2(k)$ denote the spectra of two microphone signals $x_1(t)$ and $x_2(t)$, respectively. The Global Coherence Field (GCF) is obtained by evaluating the GCC-PHAT function at different arbitrary positions in the room and averaging the values for $|\mathcal{Q}|$ different microphone pairs $(i,h) \in \mathcal{Q}$, where $\mathcal{Q}$ is the set of the microphone pairs of array $m$. It can be expressed as

$$\mathrm{GCF}_m(\boldsymbol{p}) = \frac{1}{|\mathcal{Q}|} \sum_{(i,q)\in\mathcal{Q}} R_{ih}(\delta_{ih}(\boldsymbol{p})) \ , \qquad (2)$$

where $\delta_{ih}(\boldsymbol{p})$ is the hypothetical time delay for the microphone pair $(i,h)$ for a source positioned at $\boldsymbol{p}$.

The OGCF is a score for each of $N$ candidate orientations, represented by the points $\boldsymbol{o}_j$, arranged on a regular angular grid. It is obtained as

$$\mathrm{OGCF}(\boldsymbol{s}, \boldsymbol{o}_j) = \sum_{m=0}^{M-1} \mathrm{GCF}_m(\boldsymbol{p}_m) w(\theta_{\boldsymbol{p}_m \boldsymbol{o}_j}) \ , \qquad (3)$$

where $\boldsymbol{p}_m$ is a hypothetical source position placed at the intersection of the line connecting the centroid of the microphone array $m$ and the speaker position $\boldsymbol{s}$ and a circle with radius $r$ centered at $\boldsymbol{s}$. The placement of the points $\boldsymbol{o}_j$ for six arbitrarily placed microphones and five arbitrary candidate orientations

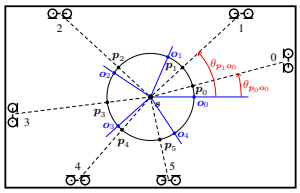

Fig. 1. A graphical representation of candidate head rotations used in OGCF [8].

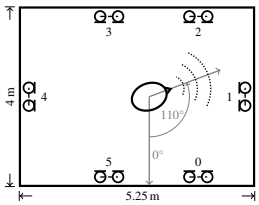


Fig. 2. The layout of the room in which the data was recorded [5].

is shown in Fig. 1. Each GCF value is weighted using the Gaussian function

$$w(\theta_{\boldsymbol{p}_m \boldsymbol{o}_j}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\theta^2_{\boldsymbol{p}_m \boldsymbol{o}_j}}{2\sigma^2}} \ , \qquad (4)$$

with the angle $\theta_{\boldsymbol{p}_m \boldsymbol{o}_j}$ between the lines connecting $\boldsymbol{p}_m$ with the speakers position $\boldsymbol{s}$ and the line connecting the evaluated orientation $\boldsymbol{o}_j$ with $\boldsymbol{s}$. The standard deviation $\sigma$ can be used to adapt the weighting pattern to the head radiation pattern.

### B. High-to-Low-Band Energy Ratio

The version of the HLBR method described in this paper was proposed in [3] as vectorial HLBR. The HLBR captures the frequency dependency and angular dependency of the sound energy radiated from the head and is defined as

$$\mathrm{HLBR}_{mq} = \frac{\sum_{k=k_{\mathrm{h,min}}}^{k_{\mathrm{h,max}}} X_{mq}(k)^2}{\sum_{k=k_{\mathrm{l,min}}}^{k_{\mathrm{l,min}}} X_{mq}(k)^2} \ , \qquad (5)$$

which is the ratio between the energy of the microphone signal $X_{mq}(k)$ in a high and a low frequency band, ranging from $k_{\mathrm{h,min}}$ to $k_{\mathrm{h,max}}$ and from $k_{\mathrm{l,min}}$ to $k_{\mathrm{l,max}}$, respectively. While the energy of the high frequency band is highly directive, the energy of the low frequency band is almost uniformly distributed in angle and is used as a normalization. This normalization for each microphone is needed as it allows the direct comparison of the HLBR values across the microphones without the need of perfect calibration of the microphones. The estimation of the head orientation from the HLBR values will be described in Section IV.

## III. Proposed Acoustic Features

Additionally to the two established features, three features are newly proposed for acoustic head orientation estimation in this work. The Averaged Coherent-to-Diffuse Power Ratio (ACDR), like the OGCF, is based on the correlation between the microphone signals. Like the HLBR, the Spectral Difference (SD) exploits the fact that the signal energy radiated to the rear of the speaker's head is lower than at the front of the speaker. The High-Band Variance (HBV) expresses the lower energy in the back as an attenuation of the higher frequencies.

### A. Coherent-to-Diffuse Power Ratio

The CDR quantifies the diffuseness of a signal using the ratio between the power of the coherent and the diffuse part of the signal. It has already been used for different purposes, such as speaker localization [11] and dereverberation [14], but has not been used for head orientation estimation before. In [14]

a DOA-independent CDR estimator using the coherence between two microphone signals and a model of the coherence of the diffuse signal component was proposed. While the coherence of the microphone signals can be calculated using the estimated Power Spectral Density (PSD) of the microphone signals, the coherence function of the diffuse part of the signal is modeled as

$$\Gamma_{\mathrm{n}}(k) = \frac{\sin(\omega_k d_{ih}/c)}{\omega_k d_{ih}/c} \ , \tag{6}$$

where $d_{ih}$ is the distance between the microphones of the pair $(i,h) \in \mathcal{Q}$ and $c$ is the speed of sound.

In order to use the CDR measure for head orientation estimation, it is computed for, and averaged over, all microphone pairs $(i,h) \in \mathcal{Q}$ of each array $m$. To support the wide range of CDR values, the natural logarithm of the average CDR value of array $m$ is taken as the feature, called Averaged CDR ($\mathrm{ACDR}_m$).

### B. High-Band Variance

The use of the HBV as a feature is motivated by the fact that the attenuation of the signal radiated to the rear of the head is higher compared to energy radiated towards the front of the head. This can be seen in a lower variance of the absolute high-frequency spectrum behind the head. The variance of the absolute high-band spectrum, ranging from frequency bin $k_{\mathrm{min}}$ to $k_{\mathrm{max}}$, is computed for all microphones

$$\sigma_{mq} = \frac{1}{k_{\mathrm{max}} - k_{\mathrm{min}} + 1} \sum_{k=k_{\mathrm{min}}}^{k_{\mathrm{max}}} \left(|X_{mq}(k)| - \mu_{mq}\right)^2 \ , \tag{7}$$

where $\mu_{mq}$ is the mean of the absolute spectrum $|X_{mq}(k)|$ in the used frequency band. In a second step, the single variances are summed over all microphones of each array

$$\mathrm{HBV}_m = \frac{1}{d_{\boldsymbol{s}m}} \sum_{q=0}^{Q-1} \sigma_{mq} \ , \tag{8}$$

with the distance $d_{\boldsymbol{s}m}$ between the speaker $\boldsymbol{s}$ and the centroid of the array $m$. The normalization has been added to assign a larger weight to the microphones closer to the speaker than to those with a larger distance, as the sound field at microphones closer to the speaker is less diffuse.

### C. Spectral Difference

The SD exploits the fact that the magnitude spectrum of the microphone signal in front of the talker is on average higher than at the back. This difference is especially large in the high-frequency band. In order to decrease the influence of noise and the speech content on this measure, the spectrum has been time-averaged, cepstrally liftered and averaged over all microphones of each array. The resulting absolute spectrum is denoted $|X'_m(k)|$. The average of all absolute spectra is calculated as

$$\bar{X}(k) = \frac{1}{M} \sum_{m=0}^{M-1} |X'_m(k)| \tag{9}$$

and the SD is defined as the frequency-averaged difference

$$\mathrm{SD}_m = \frac{1}{k_{\mathrm{max}} - k_{\mathrm{min}} + 1} \sum_{k=k_{\mathrm{min}}}^{k_{\mathrm{max}}} \left(|X'_m(k)| - \bar{X}(k)\right) \ . \tag{10}$$

## IV. Head Orientation Estimation

The head orientation estimation is based on the features presented in Section II and in Section III. In Section IV-A the methods for the orientation estimation using individual features are introduced and in Section IV-B two algorithms for the orientation estimation based on multiple features are described.

### A. Individual Features

The estimation of head orientation using the OGCF is made by a maximization. As a high value of the OGCF corresponds to a high correlation of the microphone signals coming from the assumed direction and the speech signal behind the head is more diffuse than in front of the head, the orientation for which the OGCF is maximum is taken as source orientation. The number of candidate orientations can be chosen during the calculation of the OGCF.

For the other features the number of orientations cannot be chosen during their calculation and therefore a different method has been used for the orientation estimation. The used method is part of the HLBR implementation proposed by [3]. For each microphone array $m$, a vector $\boldsymbol{v}_m$ pointing from the source position $\boldsymbol{s}$ to the centroid of the array is calculated, normalized and weighted with the feature value of the respective array. All these vectors are then summed and the orientation $o$ can be estimated as

$$\hat{o} = \angle \boldsymbol{v}_{\mathrm{sum}} \quad \text{with} \quad \boldsymbol{v}_{\mathrm{sum}} = \sum_{m=0}^{M-1} \frac{\boldsymbol{v}_m}{||\boldsymbol{v}_m||_2} a_m \ , \tag{11}$$

where $a_m \in \{\mathrm{HLBR}_m, \mathrm{ACDR}_m, \mathrm{HBV}_m, \mathrm{SD}_m\}$ (or combinations thereof, see Section IV-B) is the feature average over all microphones of each array and $\angle$ denotes the calculation of the angle with respect to a reference vector indicating an orientation of 0°. The reference vector for this work is defined in Fig. 2. This method will be referred to as Vectorial Orientation Decision (VOD).

### B. Feature Combination

In order to improve the head orientation estimation accuracy, the single features are combined using two different methods. The most straightforward way is to combine the different features linearly for each array as

$$a_m = \lambda_0 + \sum_{n=0}^{N-1} \lambda_n f_{n,m} \ , \tag{12}$$

where $f_{n,m}$ is the $n$th feature of microphone array $m$. The weights $\lambda_n$ can either be learned using a regression algorithm or can be chosen by hand. Finally, the orientation is estimated using VOD following (11). This method will be referred to as vectorial combination (VC).

TABLE I
PARAMETERS OF THE DIFFERENT FEATURES.

| | frequency band(s) | STFT length | STFT window | sequence length |
|---|---|---|---|---|
| OGCF | 0 Hz – 22050 Hz | 25 ms | von Hann | 0.5 s |
| HLBR | 200 Hz – 800 Hz<br>6500 Hz – 8500 Hz | 110 ms | von Hann | 2 s |
| ACDR | 0 Hz – 22050 Hz | 120 ms | Blackman | 2 s |
| HBV | 5000 Hz – 8000 Hz | 30 ms | von Hann | 2 s |
| SD | 5000 Hz – 8000 Hz | 60 ms | von Hann | 2 s |

As the mapping of the features to the orientation is most likely non-linear, a small artificial neural network (ANN) has been investigated as a second method. The network has one input layer, one hidden layer and one output layer. The inputs to the ANN are the feature values as an $N \cdot M$-dimensional vector. The input layer has the same number of neurons as inputs and the hidden layer has half the number of neurons as the input layer. The output layer has only a single neuron. All layers are densely connected. The activation function of the first layer is linear and a rectified linear unit is used for the hidden layer. A sigmoid function is used for the output layer. The ANN design was determined experimentally. In contrast to the VOD and the OGCF orientation estimation, the speaker and microphone positions do not need to be known for the ANN.

## V. EXPERIMENTS

### A. Data

The dataset [5] consists of speech signals recorded at a sampling frequency of 44.1 kHz of 7 male subjects in 4 different static azimuthal directions (0°, 90°, 180°, 270°). The speech is captured by 6 T-shaped microphone arrays, distributed along the walls (see Fig. 2). The room size is 3.97 m×5.25 m×4.00 m with a reverberation time of approximately 400 ms. The recorded signals have been segmented to a length of 2 s containing only speech for the evaluation of the individual features and the VC. For the training and evaluation of the ANN, segments with a length of 0.5 s have been extracted.

### B. Experimental Evaluation

Some crucial parameter values for computing the features have been selected based on a preliminary experimental analysis. Table I shows common parameter values for OGCF, HLBR, ACDR, HBV and SD. Additionally, for the OGCF algorithm $r$ has been set to 70 cm, 72 candidate orientations have been evaluated and $\sigma$ is chosen as 0.5. For the SD cepstral lowpass liftering with cutoff bin 12 of 512 has been performed. The linear combination weights of the VC were set to 1, except for $\lambda_0$ which has been set to 0.

## VI. RESULTS AND DISCUSSION

### A. Individual Features

The five different features have been evaluated on 40 segments of the dataset described in Section V-A. The error of the estimation can be seen in the blue boxplots in Fig. 3. The

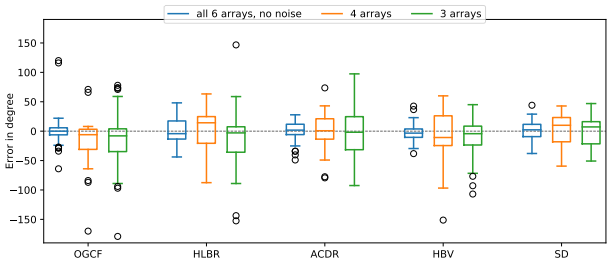

Fig. 3. The error between the estimate and the ground truth of the head orientation using different microphone arrays for the individual features.

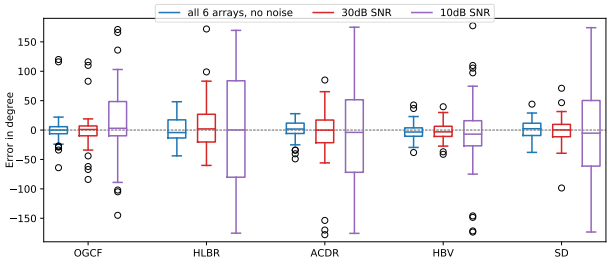

Fig. 4. The error between the estimate and the ground truth of the head orientation using different amounts of additive white Gaussian noise for the individual features.

mean value and the Root Mean Square Error (RMSE) for all experiments and methods can be found in Table II. It can be seen that the baseline algorithm OGCF performs very well for most of the trials. Most of the poor estimates of the OGCF algorithm are in the 270° direction, where the microphone array has a slightly higher distance to the speaker. The higher distance causes a lower correlation between the microphone signals of one array and leads therefore to lower GCF and OGCF values, which causes the algorithm to detect the wrong head orientation. The other methods presented here do not exhibit this degradation. The best results are achieved by the SD estimation, closely followed by the HBV.

In a next step, the robustness of the features against a changing number of microphone arrays has been investigated on a set of 3 and 4 arrays, respectively. For the 4-array test, the arrays 0, 1, 3 and 4 performed best in a preliminary evaluation, for the 3-array test the arrays 1, 3 and 5 (see Fig. 2). A further reduction of the number of microphone arrays is not feasible, as the algorithms are designed for multiple arrays distributed around the speaker. The results, depicted by in the yellow and green boxplots in Fig. 3 and in Table II, show that some algorithms are more sensitive regarding the number of microphones (OGCF, HLBR, ACDR) than others (HBV, SD). The decreasing error for the set of three arrays compared to four arrays indicate that, for HBV and SD, the positioning of the microphone arrays is more important than the number of microphones.

Finally, the sensitivity of the five features against additive white Gaussian noise has been evaluated. The noise was added to the microphone signals additionally to the background noise already included in the database. After the superposition, the ratios of the additive noise relative the dataset signal were 30 dB and 10 dB, respectively. The results in Fig. 4 and in Table II, indicate that the HBV and SD perform best with a low

TABLE II
THE MEAN ERROR AND THE RMSE FOR ALL ALGORITHMS AND EXPERIMENTS.

| | all 6 arrays, no noise | | 4 arrays | | 3 arrays | | 30 dB SNR | | 10 dB SNR | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mean error | RMSE | mean error | RMSE | mean error | RMSE | mean error | RMSE | mean error | RMSE |
| OGCF | 2.03° | 31.0° | 16.10° | 42.07° | 15.97° | 54.66° | 0.65° | 36.91° | 14.28° | **68.38°** |
| HLBR | 1.21° | 22.68° | 1.39° | 37.51° | 12.80° | 52.54° | 8.10° | 43.33° | **7.53** | 99.62° |
| ACDR | **0.32°** | 17.41° | **0.54°** | 31.15° | 2.74° | 37.12° | 8.88° | 55.0° | -9.63 | 91.32° |
| HBV | 2.61° | 16.12° | 8.79° | 44.51° | 11.36° | 34.93° | 2.67° | **16.43°** | -8.47° | 71.05° |
| SD | 1.67° | **15.03°** | 1.50° | **25.52°** | **2.50°** | **25.69°** | **0.15°** | 25.82° | -9.33° | 90.61° |
| VC | **-0.80°** | 13.70° | **-1.41°** | 25.17° | -5.92° | 22.00° | -1.00° | **14.03°** | **-2.86°** | **71.49°** |
| ANN | -2.52° | **9.09°** | -3.43° | **11.26°** | **3.84°** | **17.92°** | **-0.97°** | 15.39° | -7.82° | 78.02° |

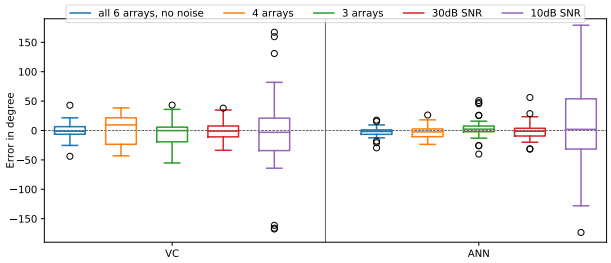

Fig. 5. The results of VC and the ANN on the five evaluation tests.

amount of noise and none of the algorithms is able to estimate the head orientation accurately at an SNR of only 10 dB.

### B. Feature Combination

For the feature combination described in Section IV-B, the same experiments as for the individual features have been performed. The VC has been evaluated on 40 segments with a length of 2 s. The training of the ANN has been performed using 200 segments of the database with a length of 0.5 s. For the evaluation 40 segments with a length of 0.5 s have been used. The results of the experiments are summarized in Fig. 5 and Table II.

It can be seen that the ANN performs best in almost all tests. Only in the case with additive white noise the VC performs slightly better. The performance of the ANN, especially in the case with an SNR of 10 dB, can most likely be enhanced by increasing the amount of training data. However, the VC shows that also in the case where no training data is available the combination of the features results in a higher accuracy relative to all individual features.

## VII. Conclusion and Future Work

In this work we proposed three new features for the head orientation estimation, which are shown to increase the estimation accuracy, as can be seen in Table II. Furthermore, the new estimation methods also allow for a higher accuracy than the known methods, if fewer microphone arrays are available. For the HBV and SD, also the robustness against a low level of noise has been increased. The combination of the five presented features showed that the accuracy and robustness of the estimation can be further increased, even without training data. In future work the robustness of the algorithms is planned to be investigated further, also with respect to a non-centered speaker position and obstruction of the direct sound path. Furthermore, the algorithms could be extended to perform position estimation and head orientation tracking.

## VIII. Acknowledgment

We would like to thank Sam Bellows and Timothy Leishmann for the access to the detailed head orientation measurements [15].

## References

[1] C. Segura, A. Abad, J. R. Casas, and J. Hernando, "Multimodal Head Orientation Towards Attention Tracking in Smartrooms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (Honolulu, USA), pp. 681–684, Apr. 2007.

[2] S. Chakrabarty, D. Pilakeezhu, and E. A. P. Habets, "Head-Orientation Compensation with Video-Informed Single Channeled Speech Enhancement," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, (Xi'an, China), Sept. 2016.

[3] A. Abad, C. Segura, C. Nadeu, and J. Hernando, "Audio-based approaches to head orientation estimation in a smart-room," in *Interspeech*, (Antwerp, Belgium), pp. 590–593, Aug. 2007.

[4] E. Murphy-Chutorian and M. M. Trivedi, "Head Pose Estimation in Computer Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[5] C. Segura and J. Hernando, "3D Joint Speaker Position and Orientation Tracking with Particle Filters," *Sensors*, pp. 2259–2279, 2014.

[6] M. Barnard and W. Wang, "Audio Head Pose Estimation using the Direct to Reverberant Speech Ratio," *Speech Communication*, 2016.

[7] J. M. Sachar and H. F. Silverman, "A Baseline Algorithm for Estimation Talker Orientation using Acoustical Data From a Large-Aperture Microphone Array," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, (Montreal, Quebec), pp. 65–68, IEEE, May 2004.

[8] A. Brutti, M. Omologo, and P. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays," in *Interspeech*, (Lisbon, Portugal), pp. 2337–2340, Sept. 2005.

[9] J. H. DiBiase, *A High-Accuracy , Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*. PhD thesis, Brown Univeristy, 2000.

[10] R. Al-Mafrachi, M. Gimm, and G. Schmidt, "A Robust Acoustic Head Orientation Estimation and Speech Enhancement for In- Car Communication Systems," in *Deutsche Jahrestagung für Akustik*, pp. 1360–1363, 2019.

[11] A. Brendel and W. Kellermann, "Learning-based Acoustic Source Localization in Acoustic Sensor Networks using the Coherent-to-Diffuse Power Ratio," in *26th European Signal Processing Conference (EUSIPCO)*, (Rome, Italy), Sept. 2018.

[12] A. C. C. Warnock, W. T. Chu, and J.-C. Guy, "Directivity of Human Talkers," *Canadian Acoustics*, vol. 30, no. 3, pp. 36–37, 2002.

[13] A. Brutti, M. Omologo, and P. Svaizer, "Estimation of talker's head orientation based on Oriented Global Coherence Field," in *Audio Engineering Society Convention*, (Paris, France), May 2006.

[14] A. Schwarz and W. Kellermann, "Coherent-to-Diffuse Power Ratio Estimation for Dereverberation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 6, pp. 1006–1018, 2015.

[15] S. D. Bellows and T. W. Leishmann, "High-Resolution Analysis of the Directivity Factor and Directivity Index Functions of Human Speech," in *Audio Engineering Society Convention*, (Dublin, Ireland), pp. 1–10, Mar. 2019.