PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Single-channel signal features for estimating microphone utility for coherent signal processing

Michael GÜNTHER, Andreas BRENDEL and Walter KELLERMANN

Multimedia Communications and Signal Processing, Friedrich-Alexander Universität Erlangen-Nürnberg,
Cauerstr. 7, D-91058 Erlangen, Germany, {michael.guenther}@fau.de

Abstract

Many microphone array signal processing techniques, e. g., for beamforming or localization, rely on coherent input signals. However, low inter-channel coherence may result from the occlusion of microphones, reverberation, or the presence of undesired signal components, so that the according signals contribute little to the overall algorithmic performance. Thus, ranking the microphone channels by their utility for subsequent coherent signal processing schemes is of considerable interest. Direct estimation of the pair-wise coherence is often straightforward in compact microphone arrays when all microphones share a common sampling clock, while wireless acoustic sensor networks require a potentially costly time synchronization of the microphone signals. In this case, estimating the channel utility ranking from simpler, single-channel features, e. g., statistical moments of the time-domain signal waveform or of the corresponding magnitude spectrum instead of the signal coherence, facilitates the clustering of useful sensor nodes for a particular task. Thereby, it is possible to determine whether it is worth the effort to synchronize the sensor signals in a sensor network, and thus save computational power and data rate if this is not the case. In this contribution, we investigate the efficacy of different single-channel signal features for determining candidate sets of sensors signals for synchronization in sensor networks.

Keywords: Microphone utility, Channel selection, Sensor network

1 INTRODUCTION

Ongoing miniaturization of microphones and advances in wireless communication allow Wireless Acoustic Sensor Networks (WASNs) to break into the consumer market. The multiple views of the acoustic scene offered by distributed sensors allow signal processing algorithms to exploit this spatial information, e. g., for spatial filtering or acoustic source localization. Many prominent examples like Minimum Variance Distortionless Response (MVDR) [9] or MUltiple SIgnal Classification (MUSIC) [7] rely, directly or indirectly, on coherent input signals. Consequently, signals with low inter-channel coherence contribute little to the overall algorithmic performance or may even be detrimental. Poor signal coherence may be caused by asynchronous sampling of channels or by a non-coherent sound field arising from, e. g., the position and orientation of the microphones relative to the acoustic source, reverberation, and the occlusion of microphones. While the former is alleviated by time synchronization of the recorded signals, the latter may require the exclusion of affected microphones from subsequent processing. If the microphone channels are already synchronized, the readily observable signal coherence offers a straightforward categorization of microphone channels by their usefulness for coherent signal processing schemes. However, the microphones in WASNs often have no common sampling clock, such that estimation of the coherence entails a potentially costly synchronization of the sensor signals. Due to limited data rate and computational power of typical network nodes, synchronization of all signals is infeasible, especially if the number of microphones is large.

Thus, in this contribution, we focus on determining candidate sets of microphones for a subsequent synchronization for coherent signal processing schemes based on established single-channel signal features. The feature values are meant to be broadcasted between network nodes to allow an informed decision which subset of channels to synchronize. Subsequently, only these candidate signals would be transmitted, such that a prescribed data rate in the WASN is not exceeded. Although synchronous sampling could reasonably be assumed for certain

subsets of sensors, e. g., for the microphones of individual compact microphone arrays in a larger network, we deliberately do not exploit inter-channel features to retain a broad scope of applicability.

In the literature, several utility measures based on inter-channel correlation have been proposed, whose common drawback is the requirement of synchronized signals. In [3], the MultiChannel Correlation Coefficient (MCCC) is employed to select a pre-specified number of channels for beamforming as preprocessing for Automatic Speech Recognition (ASR). An explicit channel utility measure for Linear Minimum Mean Square Error (LMMSE) signal estimation is proposed in [2], which relies on the availability of all sensor signals at a centralized fusion center to estimate the (inverse) signal correlation matrix. A reduction of the number of transmitted signals based on signal-subspace considerations is proposed in [1], however, still requiring the transmission of the complete signals. In [4, 6], reducing the dimensionality of observations by linear mappings is proposed, but the approach is tied to Minimum Mean Square Error (MMSE) parameter estimation. Besides correlation-based measures, the selection of a single best channel for ASR based on properties of the Room Impulse Response (RIR) [11] and based on signal features (energy, Signal-to-Noise Ratio (SNR), envelope variance and modulation spectrum area) [12] has also been investigated. Furthermore, in the context of ASR, decoder-based channel selection methods [8, 13] have been shown to improve recognition accuracy, but are only applicable to speech signals, and obviously, unless ASR systems are already in place, the addition of per-channel recognition introduces enormous computational overhead.

In this paper, vectors and matrices are denoted by bold-face lowercase and bold-face uppercase variables, respectively. $[\mathbf{x}]_k$ denotes the k-th element of the vector $\mathbf{x}$ and quantities pertaining to the n-th feature are indicated by the superscript $\cdot^{(n)}$.

The remainder of this paper is structured as follows: in Section 2 we discuss the set of employed single-channel signal features, Section 3 describes the proposed algorithm to generate candidate sets of sensors to synchronize. The experimental setup and the obtained results are discussed in Section 4, and Section 5 concludes the paper.

2 FEATURES

In this section, we give a brief overview of possible single-channel features. Many different acoustic features have been devised to describe sounds [5, 10]. Following the taxonomy in [5], we focus on instantaneous/short-time features which are computed for each time frame k. We further restrict the selection to features which are computed either directly from the temporal waveform (indicated in the following by the prefix 'T'), or from the corresponding spectral representation (indicated by the prefix 'S'). For broad applicability, we do not assume a particular signal model and also do not exploit perceptually motivated features. After these considerations, the list of investigated features is chosen to comprise:

- central statistical moments of the signal waveform: **'TSpread'**, **'TSkewness'**, **'TKurtosis'**, reflecting estimates of the standard deviation, the asymmetry and the flatness of the distribution of the sampled signal waveform, respectively
- Zero-Crossing Rate: **'TZCR'**
- statistical moments of the signal magnitude spectrum: **'SCentroid'**, **'SSpread'**, **'SSkewness'**, **'SKurtosis'**, reflecting estimates of the linear mean, the standard deviation, the asymmetry and the flatness of the magnitude spectrum interpreted as a Probability Density Function (PDF), respectively
- spectral shape descriptors:
  - **'SSlope'**, reflecting the slope of the linear regression function fitted to the magnitude spectrum
  - **'SFlatness'**, **'SAmpFlatness'**, reflecting the ratio of geometric and arithmetic mean of the power and magnitude spectrum, respectively
  - **'SRollOff'**, reflecting the frequency below which 90% of the signal energy is concentrated
- temporal variation of spectra:
  - **'SFlux'**, the difference quotient between the magnitude spectra of two successive signal blocks
  - **'SVariation'**, the complement of the normalized inner product between the magnitude spectra of two successive signal blocks

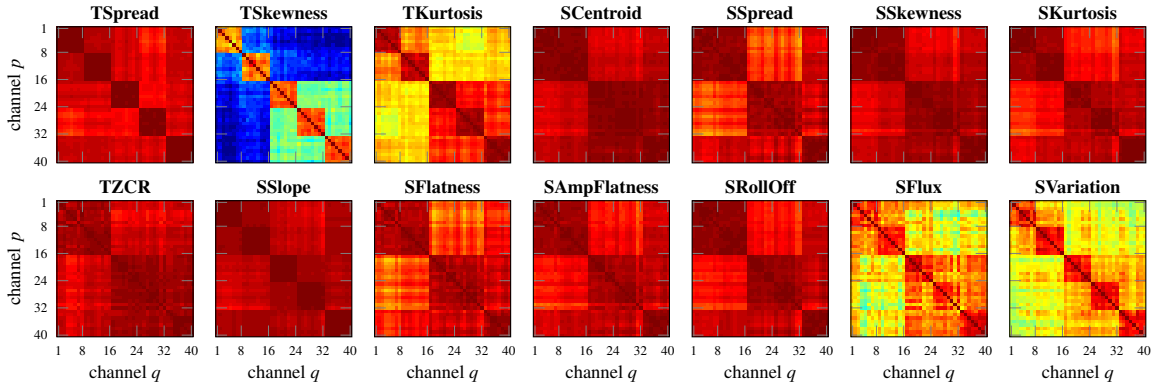


Figure 1. Correlation matrices **B** (see (5)) of different features for Experiment 1 (see Section 4.2).

For an in-depth description and mathematical formulation of the features, we refer the reader to [5].

Consider a set of P spatially distributed microphones and the corresponding signals $x_p(t)$, $p \in \{1,\dots,P\}$. Let

$$\mathbf{x}_p(k) = \begin{bmatrix} x_p(kM_p+1) & \dots & x_p(kM_p+L_p) \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{L_p} \tag{1}$$

denote the k-th signal block of L_p samples obtained by sampling the p-th microphone signal $x_p(t)$, where M_p denotes the shift between successive blocks in samples and t denotes continuous time. Note that L_p and M_p generally vary across channels due to the lack of a common sampling clock. Given a single-channel, scalar feature defined by the channel-wise mappings

$$\mathcal{F}_p : \mathbb{R}^{L_p} \to \mathbb{R}, \tag{2}$$

we collect the corresponding feature values $a_{p,k}$ from K successive signal blocks of $x_p(t)$ in the vector

$$\mathbf{a}_p = \begin{bmatrix} a_{p,1} & \dots & a_{p,K} \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^K \quad \text{with} \quad a_{p,k} = \mathcal{F}_p\left(\mathbf{x}_p(k)\right). \tag{3}$$

In the following, we omit the index p from $\mathcal{F}_p$ for the sake of a clear presentation. For the spectral features, the definition of $\mathcal{F}$ in (2) includes the required Discrete Fourier Transform (DFT).

To capture the similarity of the p-th and the q-th microphone signal w. r. t. the feature $\mathcal{F}$ while incorporating the temporal evolution of the corresponding feature values, we interpret the feature vectors $\mathbf{a}_p$, $\mathbf{a}_q$ as time series and compute the Pearson correlation coefficient

$$b_{p,q} = \frac{\sum_{k=1}^{K}\left(a_{p,k}-\overline{a}_p\right)\left(a_{q,k}-\overline{a}_q\right)}{\sqrt{\sum_{k=1}^{K}\left(a_{p,k}-\overline{a}_p\right)^2}\cdot\sqrt{\sum_{k=1}^{K}\left(a_{q,k}-\overline{a}_q\right)^2}}, \tag{4}$$

where $\overline{a}_p$ and $\overline{a}_q$ denote the arithmetic averages of the elements of the respective feature vectors. Note that $b_{p,q}$ is bounded to the range $[-1,1]$ by definition and is symmetric w. r. t. p and q. The correlation coefficients from (4) are arranged in a symmetric feature correlation matrix

$$\mathbf{B} = \begin{bmatrix} b_{1,1} & \dots & b_{1,P} \\ \vdots & & \vdots \\ b_{P,1} & \dots & b_{P,P} \end{bmatrix} \in [-1,1]^{P\times P}. \tag{5}$$

Obviously, the diagonal elements of **B** are equal to one by construction.

Recalling Section 1, the overarching goal is the categorization of microphone channels based on single-channel signal features instead of the signal coherence because the latter cannot be immediately observed due to possibly asynchronous sampling. Anticipating the in-depth description in Section 4.1, the experimental setup consists of five compact eight-channel microphone arrays. For illustration, let us consider Experiment 1 described in

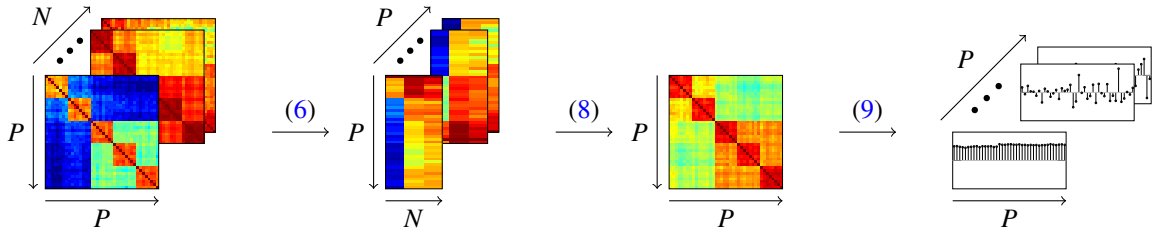


Figure 2. Schematic illustration of the proposed algorithm.

Section 4.2 as an example here, where the first two arrays are obstructed by an object. The corresponding Magnitude-Squared Coherence (MSC) matrix in Figure 4a clearly allows the identification of the obstructed arrays from their low MSC values. In line with intuition, signals of closely-spaced sensors are more coherent than those of distant sensors, reflected by larger MSC values along the blockdiagonal. To avoid costly synchronization of sensor signals, we want to extract the same channel similarity information from the feature correlation matrices, shown in Figure 1 for the same scenario. By simple visual inspection, four features stand out due to the similarity of their correlation matrices to the MSC matrix: 'TSkewness', 'TKurtosis', 'SFlux' and 'SVariation'. In the presented example, all of them reflect the obstruction of the first two arrays, as well as the higher intra-array channel similarity. Hence, we will only consider the features 'TSkewness', 'TKurtosis' and 'SFlux' in the experimental evaluation in Section 4. The feature 'SVariation' is omitted since it is conceptually closely related to 'SFlux', but does not capture the groups of obstructed and uncovered sensors as accurately.

3 ALGORITHM

In this section, we propose an algorithm to determine candidate sets of sensors for subsequent synchronization from multiple single-channel signal features, schematically illustrated in Figure 2. Interpreting correlation as an indicator for similarity of two signals, we assume that similar microphone signals also exhibit similar features. The pair-wise similarity information of microphone channels w. r. t. a set of selected single-channel features, captured by the matrix $\mathbf{B}^{(n)}$ for the n-th feature, is fused by concatenating the principal left singular vectors of the intermediate, channel-wise correlation matrices $\mathbf{C}_p$. Candidate sets of similar microphones are extracted from the resulting overall channel similarity matrix $\mathbf{R}$ by a second Singular Value Decomposition (SVD).

Given a set of N single-channel, scalar features defined by the mappings $\mathcal{F}^{(n)} : \mathbb{R}^{L_p} \to \mathbb{R}$ with $n \in \{1,\dots,N\}$, we compute the corresponding feature correlation matrices $\mathbf{B}^{(n)}$ as defined in (5). While we consider only scalar features in this contribution, in principle, vector-valued signal features may be passed to the algorithm developed in this section, if the vector elements can be treated as independent such that a feature correlation matrix can be constructed for each element.

To obtain the overall pair-wise similarity between channels across all features, we form an intermediate matrix of correlation coefficients $\mathbf{C}_p$ for each channel p w. r. t. the N selected features and compute its SVD

$$\mathbf{C}_p = \mathbf{G}_p \mathbf{D}_p \mathbf{T}_p^{\mathrm{T}} = \begin{bmatrix} b_{p,1}^{(1)} & \dots & b_{p,1}^{(N)} \\ \vdots & & \vdots \\ b_{p,P}^{(1)} & \dots & b_{p,P}^{(N)} \end{bmatrix} \in [-1,1]^{P \times N} \tag{6}$$

where $\mathbf{G}_p = \begin{bmatrix} \mathbf{g}_{p,1} & \dots & \mathbf{g}_{p,P} \end{bmatrix} \in \mathbb{R}^{P \times P}$ and $\mathbf{T}_p = \begin{bmatrix} \mathbf{t}_{p,1} & \dots & \mathbf{t}_{p,N} \end{bmatrix} \in \mathbb{R}^{N \times N}$ denote the matrix of left and right singular vectors of $\mathbf{C}_p$, respectively, and $\mathbf{D}_p \in \mathbb{R}^{P \times N}$ denotes the diagonal matrix of singular values in decreasing order. The prevalent "concept" of similarity is captured by the principal left singular vector $\mathbf{g}_{p,1}$, i. e., the first column of $\mathbf{G}_p$. By discarding the remaining singular vectors, this similarity concept is emphasized while the influence of features opposing the consensus is reduced. The principal singular vector $\mathbf{g}_{p,1}$ is normalized by

$$\mathbf{r}_p = \mathbf{g}_{p,1} / [\mathbf{g}_{p,1}]_p \tag{7}$$

such that its p-th entry is equal to one. Thus, $\mathbf{r}_p$ can be interpreted as a similarity measure between all channels and the p-th channel. Subsequent concatenation yields the final channel similarity matrix

$$\mathbf{R} = \begin{bmatrix} \mathbf{r}_1 & \dots & \mathbf{r}_P \end{bmatrix} \in \mathbb{R}^{P \times P}, \tag{8}$$

whose diagonal elements are equal to one due to the normalization in (7). However, by construction, $\mathbf{R}$ no longer exhibits the symmetry as $\mathbf{B}^{(n)}$ did, which is visible in Figure 6. We refrain from an Eigenvalue Decomposition (EVD) of $\mathbf{R}$ due to the resulting complex eigenvalues. In the special case of identical $\mathbf{B}^{(n)}$, $\mathbf{R}$ is symmetric, however no benefit is gained by considering more than one feature because $\mathbf{R} = \mathbf{B}^{(n)}, \forall n$ holds in this case. To extract candidate sets of sensors for synchronization, we perform a second SVD

$$\mathbf{R} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}, \tag{9}$$

where $\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_P \end{bmatrix} \in \mathbb{R}^{P \times P}$ and $\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \dots & \mathbf{v}_P \end{bmatrix} \in \mathbb{R}^{P \times P}$ contain the left and right singular vectors of $\mathbf{R}$, respectively, and the diagonal matrix $\mathbf{S} = \mathrm{diag}\{\sigma_1, \dots, \sigma_P\} \in \mathbb{R}^{P \times P}$ contains the corresponding singular values σ_p on the diagonal in decreasing order. Similar to the SVD in (6), the left singular vectors can be interpreted as concepts of similarity across sensors. Note that the principal left singular value $\mathbf{g}_{p,1}$ in (6) captures the dominant similarity concept of the p-th sensor to all sensors across multiple features. In contrast, the left singular vectors $\mathbf{u}_p$ in (9) reflect the similarity between *groups* of sensors by considering the square similarity matrix $\mathbf{R}$ instead of the channel-wise matrix $\mathbf{C}_p$. Thus, for each concept $\mathbf{u}_p$, a group assignment for the q-th sensor can be done by inspecting the corresponding vector element $[\mathbf{u}_p]_q$. A straightforward choice is the assignment to one of two groups of similar sensors which exhibit high intra-group similarity and low inter-group similarity based on the sign of $[\mathbf{u}_p]_q$: sensors with a positive value form one group, while the remaining sensors form the complementary group.

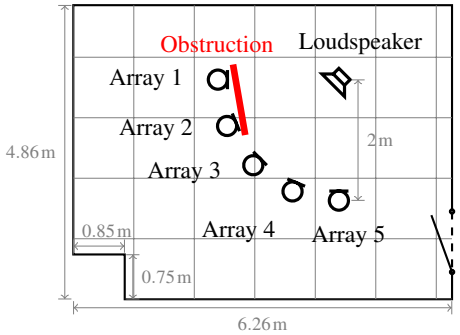
Note that there is no restriction regarding the rank of $\mathbf{R}$, such that arbitrarily many non-zero singular values can result from (9). In the extremal case of uncorrelated signal features, $\mathbf{R}$ degenerates to an identity matrix with P singular values equal to one. Choosing an appropriate number of singular values to consider is non-trivial and the discussion of such strategies is beyond the scope of this paper, although a straightforward option is basing the decision on the singular value distribution itself, e. g., by choosing all singular values above a prescribed threshold. Another option is the selection of a fixed number of promising candidate sets to be explored, depending on the available computational budget for synchronization.

4 EXPERIMENTAL EVALUATION

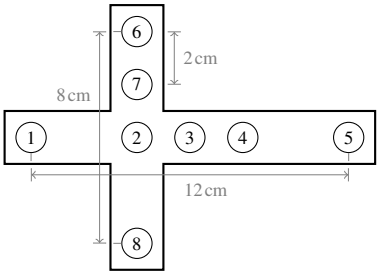
In this section, we describe the recording setup and discuss the experimental results.

4.1 Setup

The experimental setup is illustrated in Figure 3. A loudspeaker is placed in a room of dimensions $6.26\,\mathrm{m} \times 4.86\,\mathrm{m} \times 3\,\mathrm{m}$, depicted in Figure 3a, with a reverberation time of $T_{60} \approx 320\,\mathrm{ms}$. Five concentrated microphone arrays each consisting of eight microphones, depicted schematically in Figure 3b, are placed in a quarter circle around the loudspeaker at a distance of 2 m, with the loudspeaker facing the central Array 3. All microphone arrays are oriented to face the loudspeaker and are driven by a common sampling clock, which allows the signal coherence to be estimated for reference. Optionally, Arrays 1 and 2 can be obstructed by a large wooden board (roughly 1 m wide and 2 m high), shown in Figure 3a as a bold, red line. The signals are processed at a sampling rate of $f_\mathrm{s} = 16\,\mathrm{kHz}$. A signal block consists of $L_p = L = 1024$ samples (64 ms), while successive blocks are shifted by $M_p = M = 512$ samples (32 ms) for an overlap of 50%. We employ 22 s of adult male speech as the loudspeaker signal resulting in $K = 690$ available signal blocks, although our experiments have shown very similar results with adult female speech, child speech and music. As explained in Section 2, the employed set of features consists of 'TSkewness', 'TKurtosis' and 'SFlux', so that $N = 3$.



(a) Schematic illustration of the recording room. Optionally, line-of-sight from the loudspeaker to Array 1 and Array 2 can be blocked by a wooden board (1 m wide, 2 m high).



(b) Schematic illustration of the employed microphone array (frontal view). Each arrays consists of eight microphones arranged in a cross shape.

Figure 3. Experimental setup.

4.2 Experiment 1: with obstruction

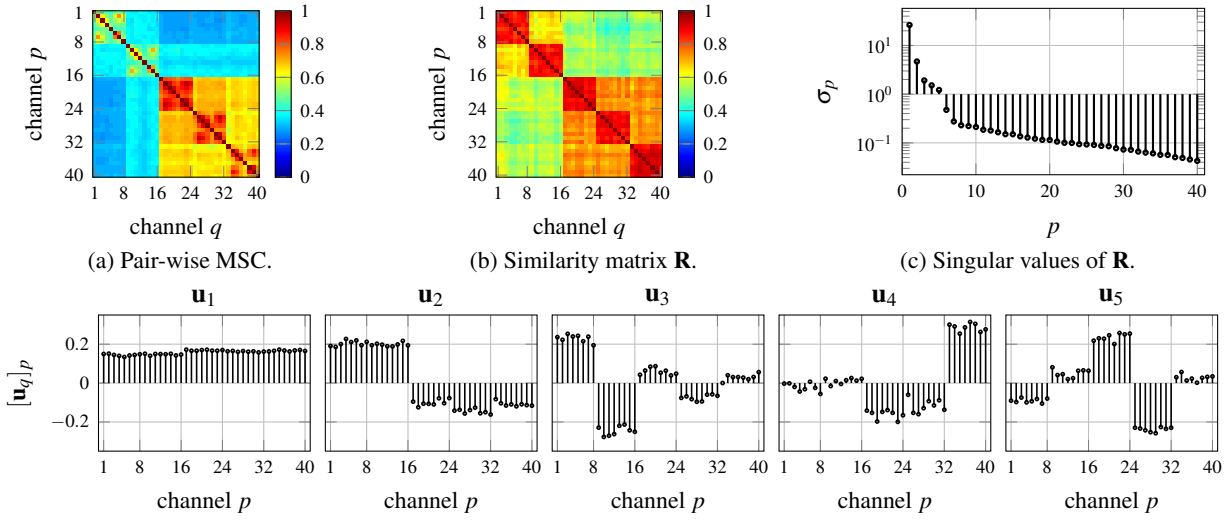
For the first experiment, we consider the setup in Figure 3a with the obstruction present, i. e., microphones 1–16 are obstructed while microphones 17–40 are uncovered. While the obstruction does not affect the speech intelligibility when listening to the signals, a clear reduction of the MSC is visible in Figure 4a which is reflected by the similarity matrix $\mathbf{R}$ in Figure 4b. The corresponding singular values and left singular vectors are shown in Figure 4c and Figure 4d, respectively. The first singular vector, corresponding to the dominant singular value, highlights the overall similarity of all channels in accordance with the subjective listening impression, although differences between the obstructed and uncovered microphones are visible. The second singular vector, whose singular value is still noticeably larger than the remaining ones, emphasizes the difference between the group of obstructed microphones (channels 1–16) and the uncovered microphones (channels 17–40). The third singular vector captures the differences between the two obstructed arrays (channels 1–8 vs. channels 9–16). Finally, the fourth and fifth singular vector similarly reflect the nuanced differences within the uncovered microphones.

4.3 Experiment 2: without obstruction

For the second experiment, we consider the setup in Figure 3a without obstruction. Like for the previous experiment, $\mathbf{R}$ and the results of its SVD are illustrated in Figure 5. The first singular value is now even more dominant, while the elements of the corresponding singular vector are essentially identical, highlighting the high similarity across all channels. Since there is no obstruction to introduce a clear distinction between two groups of microphones, the remaining singular vectors highlight feature variations due to other factors. For the second and third singular vector, these can reasonably be identified with the microphone array positions in the recording room. While the former distinguishes between sensors on either end of the spatial setup, the latter approximately captures the proximity of each sensor w. r. t. the acoustic axis of the loudspeaker.

4.4 Number of feature frames

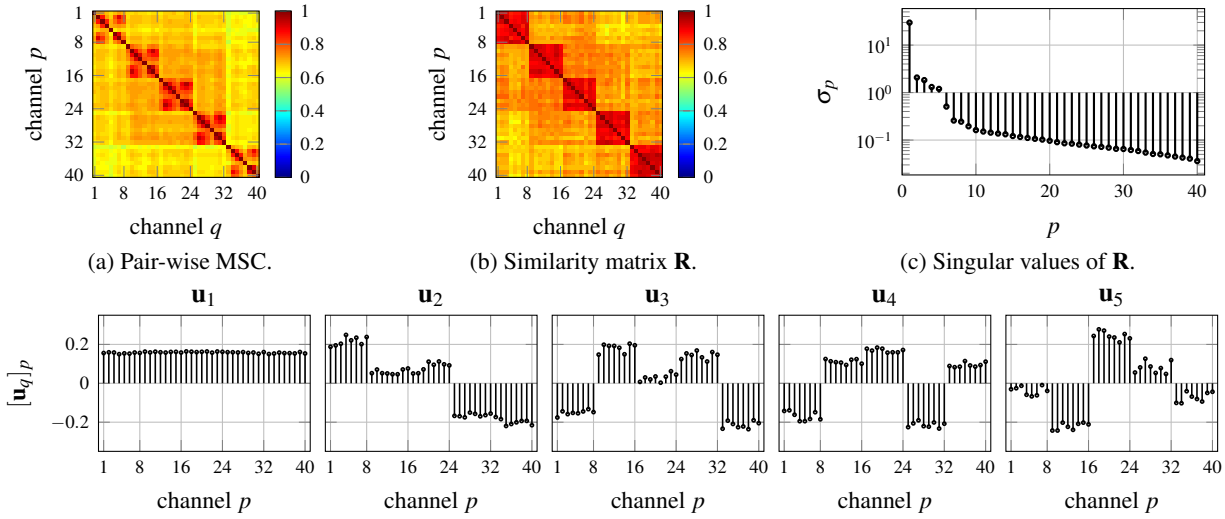
In practice, a quick decision which sensors to synchronize is obviously desirable. However, as fewer observations are used, the estimation variance of the correlation coefficient in (4) increases. Hence, we investigate the structure of $\mathbf{R}$ under the conditions of Experiment 1 for different numbers of signal blocks K by truncating the recorded signals after the specified number of blocks. The results in Figure 6 indicate that a correct categorization of channels is possible for as low as $K = 30$ (1 s signal duration). For even shorter durations like $K = 15$, a lot of the structure in columns 1–8 is lost. Note that in all cases, a signal onset period is present, during which the employed features exhibit the largest variability. In conclusion, while more signal blocks reduce the estimation variance, the presence of signal onsets appears vital for an accurate categorization.



(a) Pair-wise MSC. (b) Similarity matrix **R**. (c) Singular values of **R**.

(d) Left singular vectors of **R**.

Figure 4. Results for Experiment 1 (microphones 1–16 obstructed, microphones 17–40 uncovered.).



(a) Pair-wise MSC. (b) Similarity matrix **R**. (c) Singular values of **R**.

(d) Left singular vectors of **R**.

Figure 5. Results for Experiment 2 (all microphones uncovered).

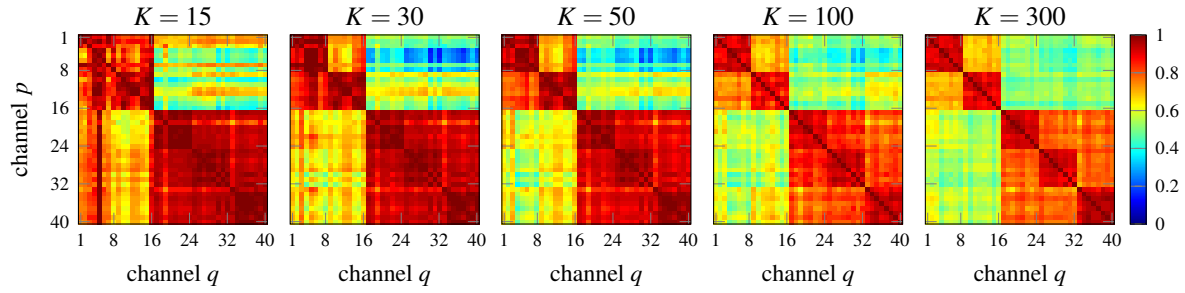


Figure 6. **R** for different number of signal blocks K.

5 CONCLUSION

In this contribution, we demonstrated the efficacy of different single-channel signal features for estimating the microphone utility for coherent signal processing. We presented a method to fuse the channel similarity information w. r. t. different features by computing the principal singular vector of an intermediate, per-channel similarity matrix. Candidate groups of sensors for subsequent synchronization are obtained by SVD of a matrix constructed from the principal singular vectors reflecting the prevalent "concept" of similarity of all channels w. r. t. a particular reference channel. The experiments confirm that groups of similar sensors can be identified for non-Gaussian signals. The presence of signal onset periods allows a categorization of sensors when only few observations are available.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] A. Bertrand and M. Moonen. Distributed adaptive estimation of correlated node-specific signals in a fully connected sensor network. In *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 2053–2056, Taipei, Taiwan, Apr. 2009.

[2] A. Bertrand and M. Moonen. Efficient sensor subset selection and link failure response for linear MMSE signal estimation in wireless sensor networks. In *European Signal Process. Conf. (EUSIPCO)*, pages 1092–1096, Aalborg, Denmark, Aug. 2010.

[3] K. Kumatani, J. McDonough, J. Lehman, and B. Raj. Channel selection based on multichannel cross-correlation coefficients for distant speech recognition. In *Joint Workshop Hands-free Speech Commun. Microphone Arrays (HSCMA)*, pages 1–6, Edinburgh, UK, May 2011.

[4] Z.-Q. Luo, G. Giannakis, and S. Zhang. Optimal linear decentralized estimation in a bandwidth constrained sensor network. In *Proc. Int. Symp. Inf. Theory (ISIT)*, pages 1441–1445, Adelaide, Australia, Sept. 2005.

[5] G. Peeters. A large set of audio features for sound description (similarity and classification). *CUIDADO project Ircam technical report*, 2004. [online] Available: http://recherche.ircam.fr/equipes/analyse-synthese/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf.

[6] I. Schizas, G. Giannakis, and Z.-Q. Luo. Distributed Estimation Using Reduced-Dimensionality Sensor Observations. *IEEE Trans. Signal Process.*, 55(8):4284–4299, Aug. 2007.

[7] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.*, 34(3):276–280, Mar. 1986.

[8] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura. Speech recognition based on space diversity using distributed multi-microphone. In *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, volume 3, pages 1747–1750, Istanbul, Turkey, June 2000.

[9] H. L. Van Trees. *Optimum array processing: Part IV of detection, estimation, and modulation theory*. John Wiley & Sons, 2004.

[10] T. Virtanen, M. Plumbley, and D. Ellis. *Computational analysis of sound scenes and events*. Springer, 2018.

[11] M. Wolf and C. Nadeu. Towards microphone selection based on room impulse response energy-related measures. In *Proc. of I Joint SIG-IL/Microsoft Workshop Speech Lang. Technol. Iberian Lang.*, page 4, Porto Salvo, Portugal, 2009.

[12] M. Wolf and C. Nadeu. On the potential of channel selection for recognition of reverberated speech with multiple microphones. In *Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Makuhari, Japan, 2010.

[13] M. Wölfel. Channel selection by class separability measures for automatic transcriptions on distant microphones. In *Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Antwerp, Belgium, 2007.