Ad Hoc Distributed Microphones Clustering: A Comparative Analysis on Using Coherence and Signal-Specific Features

Stijn Kindt[*], *Martijn Meeldijk*[*], *Nilesh Madhu* [1]

IDLab, Department of Electronics and Information Systems, Ghent University - imec, Ghent, Belgium
Email: {stijn.kindt, martijn.meeldijk, nilesh.madhu}@ugent.be

Abstract

It is often useful to cluster hoc distributed microphones according to the dominant source each captures. For example, in a recently proposed source separation approach, inter- and intra-cluster information is aggregated to enhance the dominant source at each cluster. To generate the features for this blind clustering, spectro-temporal characteristics of the signals are usually exploited to be clustered by the fuzzy C-means algorithm. A recent alternative uses the spatial relations between microphones, encoded by pairwise broadband coherence. Non-negative matrix factorisation of the positive, semi-definite coherence matrix thus obtained directly yields the cluster assignments. Here, we compare these two types of approaches in terms of the resultant cluster quality. However, coherence-based approaches require the transfer of each microphone signal to a central processing node for feature computation - in contrast to the feature-based approaches, which only require transmission of time-aggregated feature vectors. To counter this large bandwidth requirement of the coherence-based approach, we examine its performance under lossy encoding from a standard codec (BLE LC3plus). Results show that, contrary to intuition, the coherence-based approach remains robust to such non-linear encoding - making this a viable option for bandwidth-limited (wireless) acoustic sensor networks.

1 Introduction

Combining the information present in (wireless) acoustic sensor networks (WASNs), shows great potential to improve tasks such as speaker separation, diarization, and automatic speech recognition[1]. The advantage of utilising these WASNs over compact microphone arrays lies in their ability to provide additional spatial diversity. However, in the case the WASNs are constructed by ad hoc distributed microphones, their positions are not known a priori. By clustering microphones dominated by the same source, greater clarity can be achieved regarding the relative microphone positions, facilitating subsequent processing stages.

The clustering approaches discussed in [2] and [3] leverage room characteristics, utilising either diffuse noise characteristics or room impulse response (RIR) estimation. In contrast, alternative methods focus on extracting features for clustering. For instance, [4] introduced Modulation Mel Frequency Cepstral Coefficients (Mod-MFCC) as features. On the other hand, [5–7], and [8] propose the use of Deep Neural Networks (DNNs) to extract features, employing either an auto-encoder or a pre-trained speaker verification network.

Another approach exploits spatial features, specifically the magnitude squared coherence (MSC) computed between microphones [9]. These coherence values are then arranged in a coherence matrix, which is subsequently used to perform clustering via non-negative matrix factorisation (NMF) [10].

A comparison of the different techniques is currently missing. This paper aims to address this gap by comparing feature-based methods such as Mod-MFCCs and speaker embeddings with coherence-based clustering. However, in its current form, the coherence-based method requires far more bandwidth, which is usually restricted in WASNs for energy and efficiency concerns [1]. Therefore, the coherence-based method will also be evaluated with the incorporation of an audio codec. The LC3plus codec [11], used in Bluetooth low-energy (BLE)applications, can reduce a signal to a bitrate of 16kbps. Although the bandwidth required for these signals still exceeds that of the feature-based methods, it should be noted that this allocation of bandwidth is not wasted, as subsequent processing would often require access to these signals.

The rest of the paper is structured as follows: the signal model is outlined in Sec. 2, after which the coherence-based clustering is detailed in Sec. 3. This is followed by an overview of the feature-based clustering methods in Sec. 4. The experiments and results are then presented in Sec. 5 and discussed in Sec. 6. Finally, in Section 7, the paper concludes by summarising the key findings and implications, as well as providing insights into potential future research directions.

2 Signal Model

The considered scenario consists of M microphones distributed in a room with J active sources. The signal at each microphone m is given by:

$$y_m(n) = \sum_{j=1}^{J} x_{j,m}^{\text{dir}}(n) + x_{j,m}^{\text{rev}}(n) \quad + v_m(n)\,, \tag{1}$$

with n the discrete time index. $x_{j,m}^{\text{dir}}$ is the direct path contribution of source signal from the jth source to the mth microphone, while $x_{j,m}^{\text{rev}}$ represents all the reflected components of the jth source. v_m is the additive noise at the microphone m. The short-time Fourier domain representation of the signal is represented by the corresponding capital letter:

$$Y_m(l,k) = \text{STFT}\big[y_m(n)\big]\,, \tag{2}$$

where l is the time index and k is the frequency bin. The Von Hann window is used to perform the STFT.

3 Coherence-Based Clustering

First, we discuss the previously proposed frequency domain coherence-based clustering method from [9]. This approach involves calculating the magnitude squared coherence between microphone pairs and constructing a coherence matrix. Subsequently, non-negative matrix factorisation (NMF) is employed to cluster the microphones into C clusters.

3.1 Frequency-Domain Coherence

By utilising the magnitude squared coherence, it is possible to conduct an analysis of the linear relationship between two microphone signals $y_m(n)$ and $y_{m\prime}(n)$. The MSC is computed as:

$$\Gamma_{mm\prime}(k) = \frac{|P_{mm\prime}(k)|^2}{P_{mm}(k)P_{m\prime m\prime}(k)}\,, \tag{3}$$

where $P_{mm\prime}$ represents the cross power spectral density (PSD) between y_m and $y_{m\prime}$, while P_{mm} and $P_{m\prime m\prime}$ represents the auto Power spectral density of y_m and $y_{m\prime}$ respectively. The PSDs can be calculated in the STFT domain using methods such as Welch's method [12] or recursive averaging. A common approach is to

[1]This work is supported by the Research Foundation - Flanders (FWO) under grant number G081420N and imec.ICON: BLE2AV (support from VLAIO). Partners: Imec, Televic, Cochlear, and Qorvo.

[*] Equal contributions

perform averaging over 4-second sections [4]. The final coherence value is obtained by averaging over all frequency bins of the MSC:

$$C_{mm\prime} = \frac{1}{K}\sum_{k=0}^{K-1} \Gamma_{mm\prime}(k) \in [0,1]\,. \tag{4}$$

Coherence values are computed for all microphone pairs and placed in $\boldsymbol{C} \in \mathbb{R}^{M\times M}$. Note that this matrix is non-negative and symmetrical and has the property $\boldsymbol{C}_{mm} = 1$.

3.2 NMF-based Clustering

NMF [13] is employed to calculate the cluster matrix $\boldsymbol{B} \in \mathbb{R}^{M\times C}$, where $\boldsymbol{B}_{mc}$ is the contribution of microphone m to cluster c. The symmetrical and diagonal properties of the coherence matrix $\boldsymbol{C}$ can be exploited to write the problem as follows [9]:

$$\boldsymbol{C} = \boldsymbol{B}\boldsymbol{B}^T \odot (\boldsymbol{1} - \boldsymbol{I}) + \boldsymbol{I}\,, \tag{5}$$

where $\odot$ denotes element-wise (Hadamard) product, $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{1}$ is the all-ones matrix.

It is possible to estimate $\boldsymbol{B}$ using iterative multiplicative update rules based on Euclidean divergence [13]:

$$\boldsymbol{B} \leftarrow \boldsymbol{B} \odot \frac{(\boldsymbol{C} \odot (\boldsymbol{1} - \boldsymbol{I}))\boldsymbol{B}}{(\boldsymbol{B}\boldsymbol{B}^T \odot (\boldsymbol{1} - \boldsymbol{I}))\boldsymbol{B}}\,. \tag{6}$$

Due to the inherent clustering property of NMF [10], $\boldsymbol{B}$ consequently contains fuzzy membership values (FMVs) representing each microphone's contribution for each cluster. Note that these fuzzy values do not automatically sum up to one for each microphone. A normalisation step is carried out to ensure that: $\boldsymbol{B}_{mc} = \boldsymbol{B}_{mc} / \sum_{c=0}^{C-1} \boldsymbol{B}_{mc}$. These fuzzy values indicate the level to which each microphone also contains information of a different cluster and can be exploited for further separation tasks [14, 15]

4 Feature-Based Clustering

Alternative to the coherence-based method, the clusters can also be generated by comparing features extracted from the microphone signals. Ideally, these features extract the underlying dominant source and are not influenced by reverberation or interference. This paper handles the hand-crafted modulated Mel frequency cepstral coefficients (Mod-MFCC) based features proposed in [4, 16], and the speaker embeddings extracted from speaker verification (SpVer) deep neural networks, proposed in [8]. In both cases, the feature vectors are clustered using fuzzy C-means (FCM) clustering.

4.1 Mod-MFCC Features

First, the MFCCs are computed, and cepstral mean subtraction (CMS) is performed. CMS reduces the effect of reverberation and lets the features better focus on the underlying speech data [17, 18]. Taking the discrete Fourier transform (DFT) of the MFCCs, with a rectangular window of length L and modulation shift Q, generates the modulated MFCCs features. The modulation spectra are then averaged over time to account for the relatively big time shifts possible due to the inter-microphone distances in WASNs. The final features consist of two cepstral modulation ratios (CMRs) and the averaged modulation amplitude (AMA) [4, 16], each generated for 13 cepstral bins, resulting in a 39-dimensional feature vector.

4.2 Speaker Verification Features

Similar to [8], the paper takes the Emphasized Channel Attention, Propagation, and Aggregation Time Delay Neural Network (ECAPA-TDNN) [19] to generate speaker embeddings which, here, are used as clustering features. The network takes signals of arbitrary length and generates 192-dimensional speaker embeddings. ECAPA-TDNN extends the popular X-vector [20] system with an attentive statistics pooling layer, a speech-adapted version of Squeeze-Excitation (SE) [21] and multi-layer feature aggregation.

4.3 Fuzzy C-Means (FCM) Clustering

In contrast to the coherence-based method, both feature-based methods use the FCM algorithm [22] to cluster the microphones. Again $C = J+1$ clusters will be generated. The algorithm makes use of cluster centers, $\mathscr{C}_c$, and fuzzy membership values (FMV), $\mu_{m,c}$. These FMV are interpreted similarly to those at the output of the NMF. The following loss is minimised in order to come to the final clusters :

$$\mathscr{L} = \sum_{c=0}^{C-1}\sum_{m=0}^{M-1} \mu_{m,c}^{\alpha}\delta(\mathscr{F}_m, \mathscr{C}_c)\,, \tag{7}$$

where $\mathscr{F}_m$ is the feature (either Mod-MFCC or SpVer), $\delta(\mathscr{F}_m, \mathscr{C}_c)$ is the distance between $\mathscr{F}_m$ and $\mathscr{C}_c$, and α is the fuzzy weighting exponent. In contrast to the previous work, we take the cosine distance: $\delta(\mathscr{F}_m, \mathscr{C}_c) = 1 - \frac{\mathscr{F}_m^T \mathscr{C}_c}{||\mathscr{F}_m||_2 ||\mathscr{C}_c||_2}$, with $\|.\|_2$ is the ℓ_2 norm of a vector. We found that the cosine distance outperforms the Euclidean for both feature types.

5 Evaluation and Results

To evaluate the performance of the different clustering methods, experiments similar to [6] were conducted: dry speech signals of 4s length are selected from the LibriSpeech clean speech database [23]. Subsequently, $M = 16$ microphones were placed in the SINS simulated living area [24] with $J = 2$ simultaneously active speakers. The SINS database consists of realistic reverberant room impulse responses (RIRs) of an apartment. The microphone placement is constrained so that at least 3 microphones lie within the critical distance of each source, as done in [4]. This paper also limits the scenarios to cases where one source is placed at random in the left half of the room, while the second source is placed in the right half, similar to [8]. Note that with this setup, the sources can still be relatively close to each other since we do not constrain the minimal distance, but in general, they will be relatively far apart from each other. Metrics are averaged over 200 different microphone and source positions. See an example setup in Fig. 1. All signals are sampled at $f_s = 16$kHz, and the number of clusters is set to $M = J+1$, where the additional cluster represents a background (noise) cluster. Note that this is the same as done in the feature-based methods [4, 8], but different from the coherence based method, where $M = J$ was proposed to keep the comparison proper. Also, the number of sources is assumed known here, although this should normally also be estimated. However, this is out of the scope of this comparative paper.

In order to evaluate clustering performance, two classes of metrics are used as presented in [8]. The first handles the quality of the microphones in each cluster based on oracle knowledge of the underlying signals. The second looks at the quality of the subsequent cluster-based source separation proposed in [14, 15].

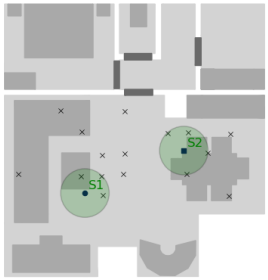


Figure 1: SINS apartment for a specific scenario. The solid dots indicate the location of the two sources, while the crosses are the microphone positions. The green circles indicate the critical distance region for each source ($d_{\text{crit}} = 0.68\,m$ for the room).

The two classes are referred to as cluster metrics and separation metrics respectively.

5.1 LC3plus Codec

Besides the performance of the different clustering methods, it is important to consider the feasibility of the methods in bandwidth-limited WASNs. The feature-based methods only require a minimal bandwidth, since they only need to transmit their features to the central node. In contrast, the coherence method requires the entire signal from each node. To alleviate this requirement, the signals can be encoded prior to transmission to the central node. While this approach still imposes a heavier burden on the network, it's not without its benefits: many other applications require access to these signals, *e.g.* for beamforming.

This paper considers the extension of the low complexity communications codec (LC3), LC3plus, at its lowest bitrate, namely 16kbps (instead of 512kbps to send 32bit samples at 16kHz). LC3plus aims to transmit high-quality audio over wireless connections at reduced bandwidth/bitrates and is used in *e.g.* Bluetooth Low Energy (BLE) and Digital Enhanced Cordless Telecommunications (DECT) [1].

To evaluate the feasibility of the coherence-based methods, we let the individual sensor nodes (each with one microphone) encode their captured signal before sending it to the central access point. There the signals will be decoded and the coherence-based clustering will be carried out on these signals.

5.2 Clustering Metrics

The distribution of the direct-to-reverberant, interference, and noise ratio (DRINR) provides insight into whether the clustering favours microphones with a strong direct-path component and a good signal-to-interference and noise ratio. DRINR is defined as:

$$\mathrm{DRINR}_{j,m} = \frac{\sum_n (x^{\mathrm{dir}}_{j,m}(n))^2}{\sum_n (y_m(n) - x^{\mathrm{dir}}_{j,m}(n))^2}, \tag{8}$$

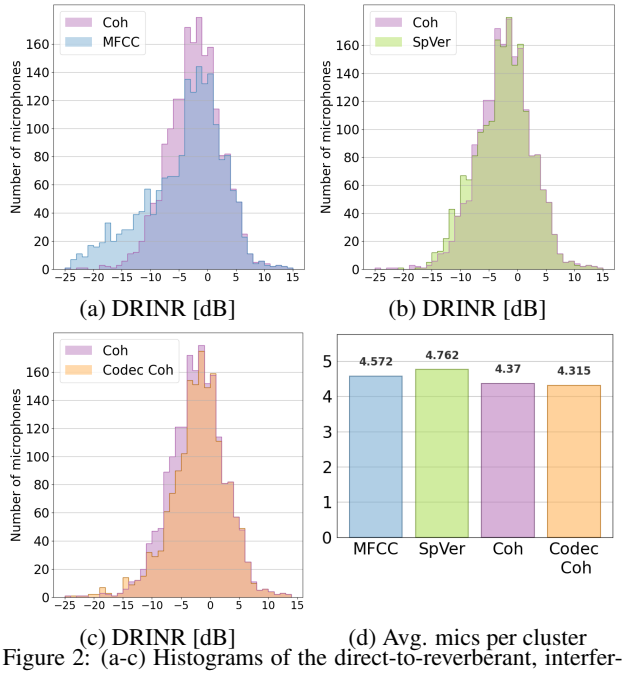
and is calculated if microphone m is part of the source cluster j. Good clusters should have many microphones with high DRINRs while avoiding including low DRINR microphones. To assess this, DRINR histograms are plotted in Fig. 2. The average number of microphones per cluster is also reported since this provides an indication of the spatial diversity within the cluster.

5.3 Separation Metrics

The main goal of clustering is to facilitate subsequent tasks such as source separation. Therefore, the quality of this separation indirectly reflects the clustering quality. Three metrics are used for this evaluation: the Source-to-Interference Ratio (SIR) [25], the Perceptual Evaluation of Speech Quality (PESQ) [26], and the Short-Time Objective Intelligibility (STOI) [27], where higher scores mean better performance.

Fig. 3 plots these metrics for the different clustering methods (colors) and separation techniques (x-axis). The four techniques are: (1) initial mask-based separation (masks), (2) delay and sum beamforming (DSB), (3) fuzzy membership value aware DSB (FMVA_DSB) and (4) a postfilter applied on the DSB (postfilter). The dotted line represents the metric in case only the reference microphone for each source is picked, showing that the separation techniques indeed improve upon selecting the best microphone. The time-frequency (TF) mask is generated by comparing the amplitude of the STFT bins between all reference microphones of each cluster. The reference microphones are determined based on the highest FMVs within each cluster. A binary TF mask is then generated by selecting those STFT bins that have a higher amplitude than those for other clusters. Here a small temporal

[1] https://www.iis.fraunhofer.de/en/ff/amm/communication/lc3.html



(a) DRINR [dB] (b) DRINR [dB]

(c) DRINR [dB] (d) Avg. mics per cluster

Figure 2: (a-c) Histograms of the direct-to-reverberant, interference, and noise ratio (DRINR). These are computed only for microphones that are part of a source cluster. (d) Average number of microphones per source cluster.

averaging is performed to account for the distances between microphones. These masks are applied to all microphones of the associated cluster, and a relative time delay is estimated, after which they are compensated for in the DSB. For the FMVA-DSB, the microphone signals contribute to the beamformed signal proportional to the FMV instead of being averaged. For the postfilter, the binary TF mask is computed w.r.t. the beamformed signals instead of the unprocessed microphone signals. For a detailed overview of these techniques, we refer to [14, 15]. Most important for this evaluation: a successful separation result requires good clusters, and can be degraded significantly with the inclusion of poor SNR microphones. Thus separation performance indirectly allows comparison of the cluster quality.

Note that for the separation, we assume that a central node has access to all the microphone signals. Even for the coherence-based method with the codec, the separation is done on the unencoded signals. This would in practice never happen, but is needed to keep the comparison fair. Speaker separation on encoded data is out of scope for this paper in order to focus on quantifying the quality of the clusters.

6 Results and discussion

6.1 Coherence- v.s. Feature-Based Clustering

Fig. 2b illustrates the distributions of the DRINRs for the coherence-based method and the SpVer features. The distribution is fairly similar, suggesting that the clustered microphones are either the same or of equivalent quality. At very high DRINRs, the distribution is even identical. However, there are still slight variations in their distributions: around -20 dB DRINR, the coherence-based method picks up slightly more outliers. In contrast, the SpVer method includes more microphones within the range of -15dB and -10dB DRINR, which are possibly not the most useful microphones, depending on how close the interferer is. From -8dB to 0dB DRINR, the coherence-based method includes more microphones with moderately good DRINRs. Additionally, Fig. 2d shows that the speaker embedding method includes more microphones on average. This, combined with the skew towards lower DRINRs suggests that the clusters from the coherence-based method should be preferred by a slight margin

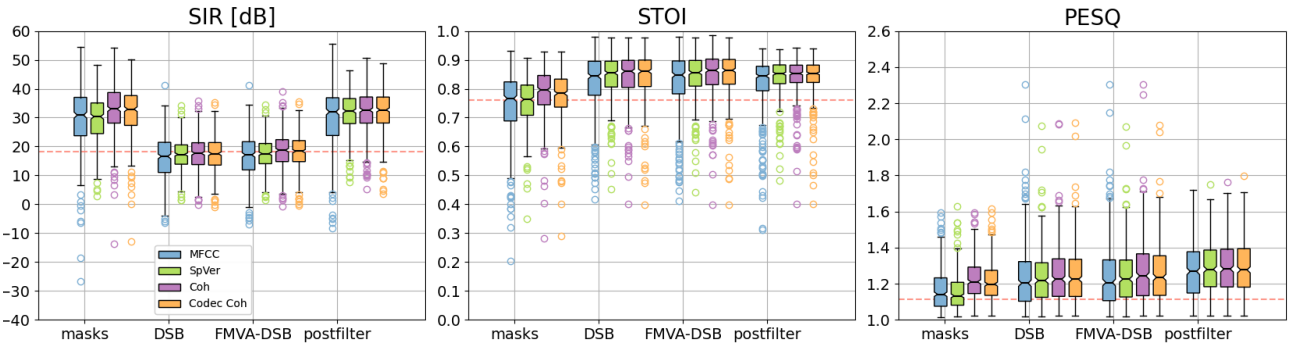


Figure 3: SIR[dB], PESQ and STOI for, Mod-MFCC and SpVer features, and coherence and coherence after LC3plus processing. The red dashed lines denote the mean of the optimal (oracle) microphones, showing optimal unprocessed performance.

for these scenarios.

If we look at the initial mask-based speaker separation metrics in Fig. 3, we notice that the coherence-based clusters clearly outperform the feature-based methods. This can be explained by the superior reference microphone selection by the coherence and NMF combination compared to the feature and FCM combination. Indeed, FCM tries to find the cluster centre that best resembles the average signal of that cluster, giving the highest FMV to the microphone closest to that cluster centre. In contrast, the coherence of all microphones of the same cluster will be highest on average towards the microphone closest to the source, resulting in the highest FMV after NMF.

However, for the subsequent separation steps, the choice of reference microphone plays no role in the separation quality, as long as the SIR of the masked signals is high enough to correctly estimate the relative delays between microphones. The gap between the feature- and coherence-based methods lessens. Nevertheless, the coherence-based method still slightly outperforms the SpVer features and clearly outperforms the Mod-MFCC features.

Furthermore, the distribution in Fig. 2a indicates that the coherence-based method clearly outperforms the method based on MFCCs, since it picks up significantly more microphones with high DRINRs and fewer with lower DRINRs. This is in line with the findings in [8], where a shoe-box simulated room was used.

6.2 Effect of LC3plus Lossy Encoding

The objective here was to determine the extent to which coherence-based cluster quality deteriorates under the non-linear, lossy encoding of the LC3plus codec at 16kbps. The results of the DRINRs (see Fig. 2c) show a remarkable resilience of the coherence-based method when confronted with LC3plus encoded signals. The encoded signals do deliver some extra outliers below -15dB DRINR, and include fewer microphones from the range -10dB to 0dB DRINR. Although the former is not a good property, the extent to which this happens is rather limited. The latter however is hard to judge and it is unclear which clustering is superior. On one hand, including more microphones increases the spatial diversity, on the other hand, it also reduces the average SNR of the clustered microphones. Fig. 2d also confirms that the unencoded signals deliver slightly larger clusters on average.

This is somewhat unexpected since coherence looks for linear dependencies between the microphone signals, and LC3plus encodes the individual signals in a lossy and non-linear way. This suggests that LC3plus encodes the signals in similar ways for each microphone signal, thus keeping same the linear dependency between the signals present before encoding.

The separation metrics in Fig. 3 illustrate, similarly to the DRINR distribution, minimal performance degradation. All metrics show that the performance degrades minimally. Only the initial masks-based separation degrades more than the other separation methods, indicating that the encoded clustering cannot recognise the optimal reference microphone as well as the unencoded clustering. However, this still exceeds the performance of the fuzzy C-means clustered signal-specific features.

Although this paper only reports the results of the lowest bitrate available in LC3plus, experiments with higher bitrates also do not degrade the clusters significantly and thus keep the linear dependencies between signals. This is not surprising, since higher bitrates require less lossy encoding.

This result establishes the coherence-base method as a viable option in bandwidth-limited WASNs with the help of a codec, requiring minimal computational power at each node and making all the individual microphone signals available for subsequent processing. Note that the optimal solution is still very application-specific, considering factors such as bandwidth and processing power requirements. For example, the lower bandwidth requirements of the feature-based clusters can be exploited by first identifying which microphone signals are needed for further processing, before sending those to the central node. *e.g.* the speaker verification only needs 192 features, which with a 32-bit representation results in ~6kb that needs to be sent to the central node per microphone. In contrast, a 4-second segment at 16kpbs still requires 64kb.

7 Conclusions

Unveiled by the cluster metrics and speaker separation metrics, the coherence-based clustering method has demonstrated a slight edge over the speaker verification method. Similarly, the evaluation confirmed that Mod-MFCC features perform worse than SpVer features and thus also the coherence-based method.

A possible downside to the coherence-based clustering method for bandwidth-limited WASNs is that all audio signals need to be transmitted to a central node, while the feature-based methods only need to share their feature vectors with a central node in order to perform the clustering. Therefore, the coherence-based method was evaluated on signals that were encoded and decoded with the reduced-bitrate codec LC3plus.

The evaluation metrics of the coherence-based clustering method with LC3plus processing demonstrate that the method remains robust in the presence of this lossy encoding. This suggests that, even though the signals are individually encoded in a non-linear manner, the linear relations between different signals are still preserved. Therefore, LC3plus encoding only slightly changes the quality of clustering. If in addition, the subsequent processing task needs access to the microphone signals, the signal transmission is far from wasted, making the encoded coherence-based method a viable option in these bandwidth-limited systems. Nevertheless, the feature-based methods could make a first selection of which microphone signals are important to be fully sent to the central node, avoiding redundant or unnecessary signals, and thereby further optimising bandwidth usage.

Further, we found that the current method to select the reference microphone from the output of the FCM algorithm, as reported in previous work, is imperfect and should be optimised in future work. Additionally, future research could evaluate the quality of source separation with LC3plus-processed signals. This analysis would provide valuable insights into the minimal bitrate, and thus bandwidth requirements in WASNs for sufficient audio quality. Lastly, more clustering techniques need to be included in the comparative study.

References

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: A signal processing perspective," in *2011 18th IEEE symposium on communications and vehicular technology in the Benelux (SCVT)*, pp. 1–6, IEEE, 2011.

[2] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 661–676, 2010.

[3] S. Pasha, Y. X. Zou, and C. Ritz, "Forming ad-hoc microphone arrays through clustering of acoustic room impulse responses," in *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, pp. 84–88, IEEE, 2015.

[4] S. Gergen, A. Nagathil, and R. Martin, "Classification of reverberant audio signals using clustered ad hoc distributed microphones," *Signal Processing*, vol. 107, pp. 21–32, 2015.

[5] A. Nelus, R. Glitza, and R. Martin, "Estimation of microphone clusters in acoustic sensor networks using unsupervised federated learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 761–765, IEEE, 2021.

[6] A. Nelus, R. Glitza, and R. Martin, "Unsupervised clustered federated learning in complex multi-source acoustic environments," in *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 1115–1119, IEEE, 2021.

[7] L. Becker, A. Nelus, R. Glitza, and R. Martin, "Accelerated unsupervised clustering in acoustic sensor networks using federated learning and a variational autoencoder," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, IEEE, 2022.

[8] S. Kindt, J. Thienpondt, and N. Madhu, "Exploiting speaker embeddings for improved microphone clustering and speech separation in ad-hoc microphone arrays," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[9] A. J. Muñoz-Montoro, P. Vera-Candeas, and M. G. Christensen, "A coherence-based clustering method for multichannel speech enhancement in wireless acoustic sensor networks," in *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 1130–1134, IEEE, 2021.

[10] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proceedings of the 2005 SIAM international conference on data mining*, pp. 606–610, SIAM, 2005.

[11] M. Schnell, E. Ravelli, J. Büthe, M. Schlegel, A. Tomasek, A. Tschekalinskij, J. Svedberg, and M. Sehlstedt, "Lc3 and lc3plus: The new audio transmission standards for wireless communication," in *Audio Engineering Society Convention 150*, Audio Engineering Society, 2021.

[12] P. Welch, "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[13] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems* (T. Leen, T. Dietterich, and V. Tresp, eds.), vol. 13, MIT Press, 2000.

[14] S. Gergen, R. Martin, and N. Madhu, "Source separation by feature-based clustering of microphones in ad hoc arrays," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 530–534, IEEE, 2018.

[15] S. Gergen, R. Martin, and N. Madhu, "Source separation by fuzzy-membership value aware beamforming and masking in ad hoc arrays," in *Speech Communication; 13th ITG-Symposium*, pp. 1–5, VDE, 2018.

[16] S. Gergen and R. Martin, "Estimating source dominated microphone clusters in ad-hoc microphone arrays by fuzzy clustering in the feature space," in *Speech Communication; 12. ITG Symposium*, pp. 1–5, VDE, 2016.

[17] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, pp. 1–22, 2004.

[18] P. N. Garner, "Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition," *Speech Communication*, vol. 53, no. 8, pp. 991–1001, 2011.

[19] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech2020*, pp. 3830–3834, International Speech Communication Association (ISCA), 2020.

[20] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[22] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & geosciences*, vol. 10, no. 2-3, pp. 191–203, 1984.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210, IEEE, 2015.

[24] R. Glitza, L. Becker, A. Nelus, and R. Martin, "Database of simulated room impulse responses for acoustic sensor networks deployed in complex multi-source acoustic environments.," in *EUSIPCO*, 2023.

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[26] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE Intl. Conf. on acoustics, speech, and signal processing.*, vol. 2, pp. 749–752, 2001.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE Intl. Conf. on acoustics, speech and signal processing*, pp. 4214–4217, 2010.