EXPLOITING SPEAKER EMBEDDINGS FOR IMPROVED MICROPHONE CLUSTERING AND SPEECH SEPARATION IN AD-HOC MICROPHONE ARRAYS

Stijn Kindt, Jenthe Thienpondt, Nilesh Madhu

IDLab, Department of Electronics and Information Systems, Ghent University - imec, Ghent, Belgium
stijn.kindt@ugent.be, jenthe.thienpondt@ugent.be, nilesh.madhu@ugent.be

ABSTRACT

For separating sources captured by *ad hoc* distributed microphones a key first step is assigning the microphones to the appropriate source-dominated clusters. The features used for such (blind) clustering are based on a fixed length embedding of the audio signals in a high-dimensional *latent* space. In previous work, the embedding was hand-engineered from the Mel frequency cepstral coefficients and their modulation-spectra. This paper argues that embedding frameworks designed explicitly for the purpose of reliably discriminating between speakers would produce more appropriate features. We propose features generated by the state-of-the-art ECAPA-TDNN speaker verification model for the clustering. We benchmark these features in terms of the subsequent signal enhancement as well as on the quality of the clustering where, further, we introduce 3 intuitive metrics for the latter. Results indicate that in contrast to the hand-engineered features, the ECAPA-TDNN-based features lead to more logical clusters and better performance in the subsequent enhancement stages - thus validating our hypothesis.

Index Terms— acoustic sensor networks, fuzzy C-means clustering, clustering metric, ECAPA-TDNN, speaker embeddings

1. INTRODUCTION

Signal capture using *ad hoc* distributed microphones, or acoustic sensor networks (ASNs), is an active and rapidly expanding field of research. With the inclusion of microphones in an increasing variety of smart devices, distributed audio capture is becoming increasingly available - with potential for application in a wide range of fields such as surveillance for assisted living and healthcare, hearing aids, communications, etc. [1]. The challenges, however, are also manifold. Compared to traditional, compact microphone arrays with predefined geometries, the relative locations of sensors are not known *a priori*, and their placement with respect to audio sources of interest can be arbitrary. The processing power and bandwidth available to each node can also be limited - constraining on-edge processing and data communication with a central hub.

In such scenarios, the challenge lies in making the optimal use of the microphones, requiring a change from the classical paradigm for compact microphone arrays. For speech separation with *ad hoc* arrays, the *microphones*, scattered across the room, are grouped into distinct clusters, where each cluster is *dominated* by a source of interest. Subsequently, by pooling the information across clusters, good first estimates of the target signals of each cluster can be obtained. These initial estimates can then drive a *within cluster* processing, to yield significantly enhanced targets.

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.

1.1. Connection to Prior Work

A multitude of different strategies have been proposed for blind clustering. In [2], for example, the magnitude squared coherence (MSC) is first computed for different microphone pairs. Subsequently, a non-negative matrix factorisation (NMF) is performed on this matrix to yield the clusters. The MSC is used in [3] too, but computed on the noise-only parts. Under the assumption of a diffuse noise field, the MSC values can then be used to estimate the distances between various microphone pairs, which is used to construct clusters based on physical distancing. In [4], the room impulse responses (RIRs) are first estimated, and then used for clustering.

In contrast to features related to the physical location of the microphones, [5, 6], proposed to compute *signal-dependent* features based on modulated Mel-frequency cepstral coefficients (Mod-MFCC). The underlying hypothesis was that microphones dominated by a source would have similar signal features – introducing a stronger dependency within the clusters, compared to just the location. In [7, 8], a simple *separation* framework based on this blind clustering was proposed, which nevertheless conclusively demonstrated the feasibility of source separation using *ad hoc* arrays with no prior knowledge of source or microphone locations.

Recently, architectures yielding deep speaker embeddings have demonstrated impressive results in speaker verification tasks and showed substantial robustness towards noise and reverberation [9]. To achieve such high discrimination, the embedding for a given speaker must necessarily be sufficiently unique. We hypothesise that such deep embeddings may be more robust indicators of source dominance and, therefore, serve as robust features for the clustering. To this end, we utilise the Enhanced Propagation and Aggregation Time Delay Neural Network (ECAPA-TDNN) [10], the state-of-the-art architecture for speaker verification, to generate the requisite features for the clustering. These are benchmarked against the hand-engineered features using the same separation framework of [7]. In addition, we also suggest 3 metrics that allow for an intuitive appreciation of the clustering performance - which are also considered for the benchmarking.

The paper is structured as follows: Sec. 2 introduces the signal model and features. The clustering and separation frameworks are discussed in Sec. 3 and Sec. 4. The experimental validation framework and the chosen metrics are described in Sec. 5, followed by the evaluation results. Sec. 7 concludes the paper.

2. SIGNAL MODEL AND FEATURES

We assume there are M microphones and J localised sources distributed in the room. The signal y_m of microphone m is given as:

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

where n is the discrete time index, $x_{j,m}$ is the source signal from the jth source to the mth microphone, which is split up into the direct path contribution $x_{j,m}^{\text{dir}}$ and the reflections $x_{j,m}^{\text{rev}}$. v_m represents the additive noise at the mth microphone. All processing is done on the short-time Fourier domain representation of the signal:

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

where l is the time index and k is the frequency bin.

2.1. MFCC-based Features

The hand-engineered features based on modulated Mel-frequency cepstral coefficients (Mod-MFCC) [5, 6] form the baseline. The Mod-MFCCs are the discrete Fourier transforms (DFTs) of the MFCC features, with a rectangular window of length L and modulation shift Q. In order to be robust against time shifts between the distant microphones in ASNs, the modulation spectra are averaged over time. Further, cepstral mean normalisation (CMN) is applied on the MFCCs, which reduces the effect of reverberations for cases where the room impulse response (RIR) stays constant [11].

As proposed in [5, 6] the Mod-MFCC-based feature vector $\boldsymbol{\mathcal{F}}^{\text{MFCC}}$ is constructed by stacking, respectively, the $\mathcal{N}$-dimensional sub-features obtained from two sets of cepstral modulation ratios ($\text{CRM}_{\kappa_1|\kappa_2}(\eta)$) and the $\mathcal{N}$-dimensional averaged modulation amplitude ($\text{AMA}(\eta)$). Here, η is the cepstral index and κ denotes the modulation frequency.

2.2. Features from speaker embeddings

For our goal of clustering microphones around speech sources, we hypothesise that embeddings generated by speaker verification networks lead to good clustering features. In verification tasks, embeddings are used to test if two audio utterances are spoken by the same person. Applied to our case, microphones dominated by the same speaker should, similarly, yield embeddings that are near identical.

For the generation of these embeddings, we take the recent Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network (ECAPA-TDNN) [10]. It maps audio of arbitrary length to 192 dimensional speaker embeddings. The speaker similarity score is determined by computing the cosine similarity between two speaker embeddings. If the similarity score exceeds a predetermined threshold, the utterances are accepted as coming from the same speaker. For the clustering, however, we compute the embeddings for each microphone and use them as the features, $\boldsymbol{\mathcal{F}}^{\text{SpVer}}$, for the clustering algorithm.

3. FUZZY C-MEANS CLUSTERING

We use the fuzzy C-means (FCM) algorithm to cluster the microphones by the extracted feature vectors, $\boldsymbol{\mathcal{F}}_m$. The goal is to separate the microphones in $C = J + 1$ fuzzy clusters – one cluster for each source and one background (noise) cluster. The algorithm returns cluster centres, $\boldsymbol{\mathcal{C}}_c$, and fuzzy membership values (FMV), $\mu_{m,c}$, that reflect how much microphone m belongs to cluster c. Thus, it can be used to determine the *reference* microphone for each cluster - which then plays a key role for the separation. The FCM algorithm minimises the following weighed least-squared error function [12]:

$$\mathcal{L} = \sum_{c=0}^{C-1}\sum_{m=0}^{M-1} \mu_{m,c}^{\alpha}\delta(\boldsymbol{\mathcal{F}}_m, \boldsymbol{\mathcal{C}}_c)\,, \tag{3}$$

where α is the fuzzy weighting exponent, typically $1 \leq \alpha \leq 2$ and $\delta(\boldsymbol{\mathcal{F}}_m, \boldsymbol{\mathcal{C}}_c)$ is the distance metric. Here we take the standard Euclidean distance: $\delta(\boldsymbol{\mathcal{F}}_m, \boldsymbol{\mathcal{C}}_c) = ||\boldsymbol{\mathcal{F}}_m - \boldsymbol{\mathcal{C}}_c||_2^2$ where $\|.\|_2$ is the ℓ_2 norm of a vector.

The reference microphone signal $Y_c^{\text{ref}}(l,k)$ of cluster c is obtained by selecting the microphone with the highest fuzzy value for that cluster:

$$\begin{aligned} Y_c^{\text{ref}}(l,k) = Y_m(l,k) \text{ if } \mu_{m,c} > \mu_{\bar{m},c}\,, \\ \forall \bar{m} \in \{0,\cdots,M-1\}, \bar{m} \neq m \end{aligned} \tag{4}$$

For separation, we transform the fuzzy clusters into hard partitionings. Thus a microphone m is allocated to cluster c if:

$$\mu_{m,c} > \mu_{m,\bar{c}}, \quad \forall \bar{c} \in \{0,\cdots,C-1\}, \quad \bar{c} \neq c\,, \tag{5}$$

and its corresponding signal is denoted as $y_{m,c}$.

4. CLUSTERING-BASED SOURCE SEPARATION

Once the clusters are identified, the separation framework is *identical* to that described in [7] and consists of the following steps. Firstly, we obtain an initial estimate of the target source in each cluster by means of time-frequency masking. These initial estimates are then used to time-align the microphone signals in the respective clusters. Following, a simple delay-and-sum beamforming (DSB) is applied to compute the enhanced target signal for the cluster. Of course, more sophisticated adaptive beamformers based on classical or data-driven approaches can be used to further enhance the signals, but since the goal of this paper is to validate deep embeddings as clustering features, we argue that benchmarking with DSB is sufficient.

4.1. Initial source estimation

To obtain the time-frequency mask for the initial source estimates, we assume that the localised sources are approximately disjoint in their STFT representation [13], so only one source may be dominant at any time-frequency (T-F) bin. We estimate T-F masks $\mathcal{M}_c(l,k)$ for each cluster as a binary mask, obtained according to [7, 8] as:

$$\mathcal{M}_c(l,k) = \begin{cases} 1 & |Y_c^{\text{ref}}(l,k)| > \dfrac{1}{B}\sum_{b=l-B+1}^{l} |Y_{\bar{c}}^{\text{ref}}(b,k)|, \\ & \forall \bar{c} \in \{0,\cdots,C-1\}, \bar{c} \neq c \\ 0 & \text{else}\,. \end{cases} \tag{6}$$

B is an averaging parameter, needed to reduce jitter in STFT amplitudes (and, consequently, the mask estimation) due to the fact that the inter-microphone delay for a source is non-negligible compared to the STFT length and frameshift for the distributed scenario, due to the much larger microphone spacings [7].

4.2. Mask-based delay and sum beamforming

The $\mathcal{M}_c(l,k)$ are then applied to the microphone signals of the respective cluster – yielding an initial estimate of the underlying source signal of *that* cluster. Using these estimates a *time-alignment* of the microphone signals in the cluster is performed. The delay $\widehat{\tau}_{m,c}$ to be compensated is computed with respect to the reference microphone of cluster c, and is obtained by a simple correlation analysis.

Having obtained the $\widehat{\tau}_{m,c}$, we apply this to the unprocessed microphone signals $y_{m,c}$ and average the signals – yielding the DSB output for cluster c.

$$\widehat{x}_c^{\text{DSB}}(n) = \frac{1}{M_c}\sum_{m} y_{m,c}(n - \widehat{\tau}_{m,c})\,, \tag{7}$$

where M_c is the number of microphones in cluster c.

5. EXPERIMENTAL SETUP

We simulated 100 different rooms using Pyroomacoustics [14]. Room dimensions were uniformly chosen from 6 different sizes $\in \{[5,5,3],[8,4,3],[6,5,3],[6,7,3],[3,4,3],[3,8,3]\}$m and with reverberation times (RT_{60}) uniformly sampled between 0.2s and 0.8s. Each scenario is simulated with $M = 15$ microphones and $J = 2$ equally loud speakers, selected from the PTDB-TUG speech datasets [15]. The ECAPA-TDNN network was trained on the *completely independent* VoxCeleb 1&2 dataset (audio of 7250 celebrities, from YouTube) [16]. One source position is randomly chosen in the left half of the room and the other in the right half of the room. The microphones are scattered across the room, but it is ensured that at least 3 microphones are within the critical distance of each source. The sources and microphones are coplanar in the z-dimension (height: 1.7m). Spatially diffuse noise at 10dB signal to noise ratio (SNR) is added, where the signal energy is measured at the *centre* of the room. Thus, individual SNRs at each microphone can strongly vary.

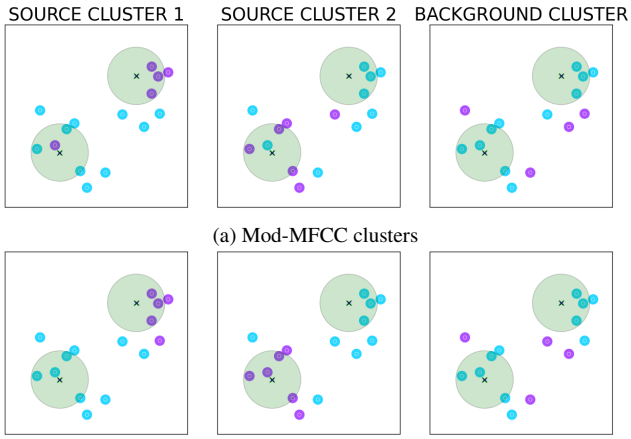
Both the MFCC-based and speaker embedding features are computed on 4-second segments of the audio signal. This segmentation is also used in the computation of the cross-correlation. All audio signals are sampled at 16kHz. For the STFT, we use the von Hann window of length 512 samples (32ms) and window shift of 160 samples (10ms). The MFCC parameters are: $L = 16$ and $Q = 8$. The number of MFCC features $\mathcal{N} = 13$, resulting in a 39-dimensional feature vector $\boldsymbol{\mathcal{F}}^{\text{MFCC}}$. Note that this is less than the dimension of $\boldsymbol{\mathcal{F}}^{\text{SpVer}}$. However, for clustering, the quality of the features is more important than the dimension of the feature vector, and increasing $\mathcal{N}$ does not necessarily result in more informative features.

The averaging factor B in (6) is set to 5. For clustering, we use the fuzzy C-means python package [17]. As the features are obtained by aggregating data across several seconds, they are inherently robust to sampling rate offsets and sampling time discrepancies between the microphones. The TDOA estimation can also compensate for asynchronous sampling, so we expect the beamforming to be minimally affected by the asynchronicity – and in a *comparable* manner for either set of features. Since this does not provide any additional indication of the quality of the clusters obtained using the Mod-MFCC or the embedding-based features, we consider the microphones to be synchronously sampled. We note that compensating for sampling discrepancies is important in practical systems and robust methods for these have been previously proposed *e.g.* [18, 19].

5.1. Evaluation metrics

Benchmarking the quality of the clustering is an open problem. Since there is no easily definable ground truth (what decides the 'goodness' of a cluster?), it is difficult to define appropriate performance metrics. Ground truths *e.g.* based on oracle knowledge of the RIRs [2] or distances [3, 4] do not convey the full picture regarding the signal mixing for the *ad hoc* scenario. The same disadvantage exists for the normalized cluster-centroid–to–source distance metric used in [6]. Further, as the position of the cluster centroid is computed based on the location of the microphones in that cluster, a single outlier can bias the results.

We suggest 3 alternative metrics which yield a more intuitive insight on the clustering performance. The first two are, respectively, (i) the *distribution* of the direct-to-reverberant ratio (DRR) and (ii) the *distribution* of the direct-to-reverberant, interference, and noise ratio (DRINR), computed for microphones allocated to *speech-source* clusters. Thus, for a microphone signal y_m, assigned to a *speech* cluster c these are defined as:



(a) Mod-MFCC clusters

(b) SpVer clusters

Fig. 1: Example clusters with both features. The room has dimensions $[5,5,3]$ m, the two crosses are the source positions and the coloured dots are the microphone positions. Each column shows a different hard cluster, where dark purple indicates that a microphone is part of the cluster.

$$\text{DRINR} = \frac{\sum_n (x_{c,m}^{\text{dir}}(n))^2}{\sum_n (y_m(n) - x_{c,m}^{\text{dir}}(n))^2} \quad \text{and} \tag{8}$$

$$\text{DRR} = \frac{\sum_n (x_{c,m}^{\text{dir}}(n))^2}{\sum_n (x_{c,m}^{\text{rev}}(n))^2}\,. \tag{9}$$

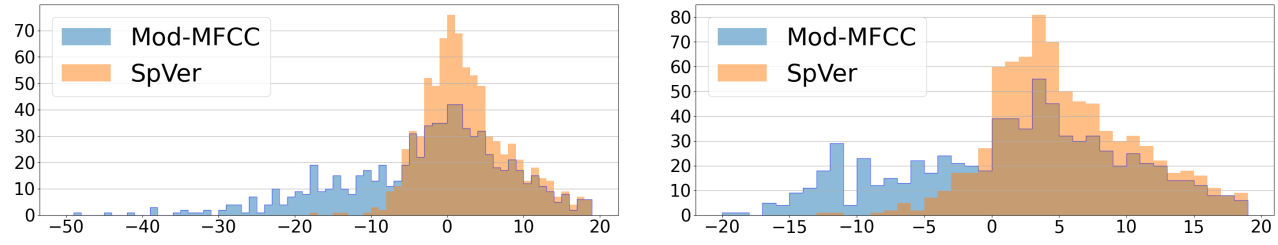
A distribution centered around high values of DRRs and DRINRs then indicates that the clustering favours microphones with a strong direct-path component and a good signal to interference and noise ratio – which is optimal for the subsequent enhancement stages. The third metric is the average number of microphones allocated to a speech cluster. In combination with the previous metrics, this indicates the amount of spatial diversity that can be exploited in extracting the desired source.

For benchmarking the features in terms of the resultant speaker separation ability we consider 3 standard instrumental metrics: the first is the source-to-interference ratio (SIR), as defined by [20]. This is most important to the masks since the masked signals are used to estimate the TDOA. However, a decent SIR can be achieved by suppressing the interfering source completely, while only keeping a small portion of the target speech. This would however lead to unintelligible, poor-quality speech. Therefore we also use perceptual quality (PESQ: perceptual evaluation of speech quality [21]) and intelligibility (STOI: short-time objective intelligibility [22]) metrics.

6. RESULTS AND DISCUSSION

6.1. Quality of the clustering

Fig. 1 depicts the clustering obtained for the two considered feature sets, for a single test condition. This allows for a visual appreciation of the quality of clustering. The speaker embedding features make rather logical clusters, while the Mod-MFCC features fail to distinguish between the two speakers. Further, for the embedding-based features, all microphones that are within the critical distance are part of their respective cluster. These features also yield a slightly larger cluster around source 1, with an extra microphone that is outside the critical distance, but is intuitively dominated by the source. Audio samples are also available at https://aspireugent.github.io/Ad-Hoc-Distributed-Microphone-Clusters/



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

Fig. 2: Histograms of the direct-to-reverberant, interference, and noise ratio (DRINR) in (a) and direct-to-reverberant ratio (DRR) in (b). These are computed only for microphones that are part of a source cluster.

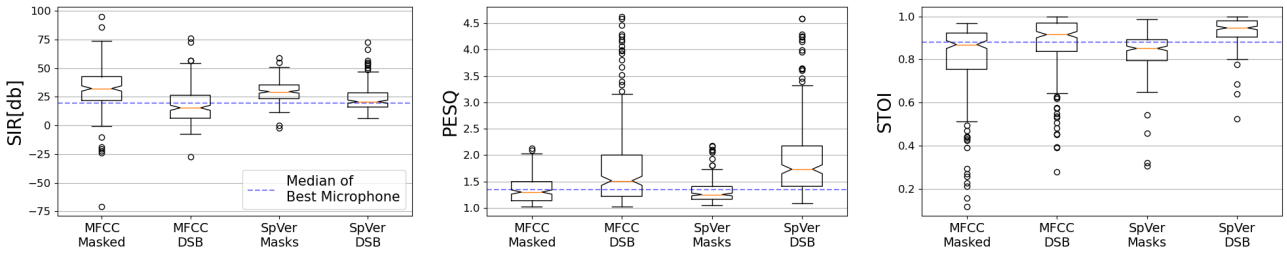


Fig. 3: Performance metrics (SIR, PESQ, STOI) showing the separation effectiveness of the cluster features (Mod-MFCC and Speaker Verification(SpVer)) and method (Masks and DSB). The dotted blue line shows the median of the metrics, computed on the best microphones (in terms of SIR) for each source. This is selected based on oracle knowledge.

Next, the distribution of the DRRs and DRINR are plotted in Fig. 2. As can be seen, when using deep embeddings, the histograms for the DRINR and DRR metrics are biased towards the right indicating that the clustering results in assigning more microphones with high DRRs and DRINRs to the speech clusters. In contrast, the histograms when using the Mod-MFCC features show a much larger spread and are less peaky. This indicates that the Mod-MFCC feature-based clusters tend to occasionally include microphones that are dominated by other sources or background noise, which can degrade the subsequent stages. Thus, we may already conclude that the average quality of the microphone signals assigned to speech clusters when using the deep embeddings is significantly higher than when using Mod-MFCC-based features.

Also in terms of the size of the speech clusters: using Mod-MFCC features results in an average of 3.81 microphones per speech cluster. This increases to an average is 4.06 microphones per cluster when using the deep embeddings, which is roughly 6.5% more. Thus, using deep embeddings results also in increased spatial diversity. Combined with the increased average quality of the microphones assigned to the speech cluster, a better separation performance is expected when using embeddings as clustering features.

6.2. Performance on source separation

The results of the source separation are depicted in Fig. 3. We notice that the mask-based separation (on the reference microphone of the clusters) yields a high SIR. This indicates that the computed masks yield a good estimate of the underlying target signal. This is important for reliable time-delay computation. This also indicates that fuzzy values are a good method of determining the reference microphone. The SIR would also suggest that mask-based separation is better than the DSB. However, PESQ and STOI show that the masks also lead to unwanted distortions in the target's speech.

To benchmark the quality of the cluster in terms of the subsequent use for *enhancement*, we compare the Mod-MFCC-based clusters to their deep-embedding counterparts in terms of the DSB results. The notched box plots show that the deep embeddings outperform Mod-MFCC features, with statistical significance at the median level. The median SIR is 5.36dB higher and the median PESQ and STOI scores are 0.22 and 0.03 higher respectively. This improved performance also indirectly indicates the assignment of higher SINR microphones to the speech clusters when using speaker embeddings - which supports our conclusions in the previous subsection.

We also note that the SIR and STOI plots of the MFCC features show more outliers at the lower end. This is also what we empirically saw in the cluster plots: in some cases, the Mod-MFCC features are not perfect in distinguishing between a source cluster and the background cluster. This further supports the conclusions we have drawn from the DRR and DRINR distributions and reinforces the usefulness of the metrics for ad hoc clustering.

7. CONCLUSIONS

We proposed to utilise speaker embedding features, gathered from the ECAPA-TDNN speaker verification network, for clustering *ad hoc* distributed microphones. Compared to the classical Mod-MFCC features, speaker embedding features generate larger and higher quality clusters – indicating more potential for exploiting spatial diversity. This last is validated by the better results of the beamformer output on SIR, PESQ and STOI. The embedding-based features also provide more logical microphone-cluster assignments compared to the Mod-MFCC features, as shown by the outliers in the STOI results for the latter.

We also showed that the distributions of the DRR and DRINR provide useful insights into the quality of the clustering, and can help interpret results obtained by the later processing stages.

Future work includes utilising the masks (or a variant of them) to perform other beamformers, like the multichannel Wiener filter. The ECAPA-TDNN framework is currently trained on single-speaker scenarios. A next step is the incorporation of our recent work in attempting to alleviate the multi-speaker robustness in ECAPA-TDNN [23].

8. 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)*. IEEE, 2011, pp. 1–6. 1

[2] 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)*. IEEE, 2021, pp. 1130–1134. 1, 3

[3] 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. 1, 3

[4] 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)*. IEEE, 2015, pp. 84–88. 1, 3

[5] 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. 1, 2

[6] 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*. VDE, 2016, pp. 1–5. 1, 2, 3

[7] 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)*. IEEE, 2018, pp. 530–534. 1, 2

[8] 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*. VDE, 2018, pp. 1–5. 1, 2

[9] Andrew Brown, Jaesung Huh, Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxsrc 2021: The third voxceleb speaker recognition challenge," *arXiv preprint arXiv:2201.04583*, 2022. 1

[10] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech2020*. International Speech Communication Association (ISCA), 2020, pp. 3830–3834. 1, 2

[11] 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. 2

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

[13] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2002, vol. 1, pp. I–529. 2

[14] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 351–355. 3

[15] Gregor Pirker, Michael Wohlmayr, Stefan Petrik, and Franz Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011. 3

[16] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech 2018*, pp. 1086–1090, 2018. 3

[17] M. L. D. Dias, "Fuzzy c-means: An implementation of fuzzy c-means clustering algorithm.," May 2019, Available: https://git.io/fuzzy-c-means. 3

[18] S. Markovich-Golan, S. Gannot, and I. Cohen, "Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4. 3

[19] T. Gburrek, J. Schmalenstroeer, and R. Haeb-Umbach, "On synchronization of wireless acoustic sensor networks in the presence of time-varying sampling rate offsets and speaker changes," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 916–920. 3

[20] 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. 3

[21] 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.*, 2001, vol. 2, pp. 749–752. 3

[22] 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*, 2010, pp. 4214–4217. 3

[23] J. Thienpondt, N. Madhu, and K. Demuynck, "Margin-mixup: A method for robust speaker verification in multi-speaker audio," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023. 4