A compressed encoding scheme for approximate TDOA estimation

Elizabeth Vargas*, James R. Hopgood†, Keith Brown* and Kartic Subr‡
*Institute of Sensors, Signals and Systems, Heriot-Watt University
†Institute for Digital Communications, University of Edinburgh
‡Institute of Perception, Action and Behaviour, University of Edinburgh
Edinburgh, United Kingdom
Email: ev42@hw.ac.uk, james.hopgood@ed.ac.uk, k.e.brown@hw.ac.uk, k.subr@ed.ac.uk

***Abstract*—Accurate estimation of Time-Difference of Arrivals (TDOAs) is necessary to perform accurate sound source localization. The problem has traditionally been solved by using methods such as Generalized Cross-Correlation, which uses the entire signal to accurately estimate TDOAs. However, this could pose a problem in distributed sensor networks in which the amount of data that can be transmitted from each sensor to a fusion center is limited, such as in underwater scenarios or other challenging environments. Inspired by approaches from computer vision, in this paper we identify Scale-Invariant Feature Transform (SIFT) keypoints in the signal spectrogram. We perform cross-correlation on the signal using only the information available at those extracted keypoints. We test our algorithm in scenarios featuring different noise and reverberation conditions, and using different speech signals and source locations. We show that our algorithm can estimate Time-Difference of Arrivals (TDOAs) and the source location within an acceptable error range at a compression ratio of $40:1$.**

***Index Terms*—microphone arrays, time difference estimation, signal compressed encoding**

I. Introduction

The literature on estimation of Time-Difference of Arrivals (TDOAs) is rich with a variety of approaches. One of the most common methods is Generalised Cross-Correlation (GCC), which is used to find the TDOA in a microphone array [1]. Methods based on cross-correlation are classified into two groups: ones that use a pair of microphones, and ones that draw on the redundancy among the microphones in the array. The first group includes the Smoothed Coherence Transform (SCOT) [2] and Generalized Cross-Correlation Phase-Transform (GCC-PHAT) [3] techniques, which are an extension of the cross-correlation into the frequency domain using a spectral normalization parameter. The second group of methods uses a spatial correlation matrix (MCCC) to determine the TDOA values that minimize the cross-correlation between each pair of signals. The most common of these methods is MUltiple SIgnal Classification (MUSIC) [4], which uses eigenvectors to estimate the TDOA.

Estimating TDOAs across a distributed sensor network is of increasing relevance as decentralised ad-hoc devices become more and more widespread. In such situations, the sensors need to exchange information to estimate the TDOAs. For example, to estimate TDOA using GCC would require transmission of the entire signal, or at the least a down-sampled version (which will lead to temporal quantisation). In scenarios in which the communications bandwidth is limited, or in which there are constraints on the amount of data that can be transmitted, approaches based on the full signal information are not very useful. Typical scenarios include underwater sensors [5], inexpensive ad-hoc mobile networks with energy constraints [6], and cases in which a high-speed communications network is either denied or unavailable (for example, disaster zones). Simon et al. [7] have developed an algorithm that relies on event detection of the signals in order to decide which parts of the signals to transmit. The authors transmit 1.1% of the raw signal, but they limited their experiments to a single scenario under specific noise and reverberation conditions. Similarly, Fuyong et al. [8] present a compression algorithm tested using compression ratios between $4:1$ and $8:1$. Additionally, there are authors who focus on sensor networks on low-bandwidth localization in [9], [10]; however, these are active sensing methods, in that sensors may emit calibration signals.

Previous studies have used different methods that involve feature extraction from the audio signals, including music identification [11], [12] and alignment of unsynchronized meeting recordings [13]. The most popular of these is known as audio fingerprinting [14], commonly used for music identification. It uses the signal spectrogram to select spectral peaks, provided that their power spectral amplitude is above a given threshold. These peaks are grouped into pairs to form a landmark, which is indexed using a hashing function. A set of these landmarks combines to characterize a song. Audio fingerprinting is used to perform *self-localization* in an ad-hoc microphone array in [15]. The problem in this instance is to localize sensors rather than sound sources, so the sources are placed in end-fire locations (i.e. points that lie on a straight line between two microphones, excluding the points that lie between the microphones) to guarantee a maximum TDOA.

In contrast to existing work that performs peak detection based on thresholding, we propose to detect audio landmarks using the Scale-Invariant Feature Transform (SIFT), a common approach in computer vision. Although there is evidence in the literature that authors have previously used SIFT on spectrograms [16]–[18], this is the first time to the best of our

knowledge that such an approach has been applied with a focus on data compression. In this paper, we present an approach based on estimating certain specific samples of the signal to be transmitted so as to estimate the delay using GCC. We use the SIFT algorithm to extract keypoints in the spectrogram, which is treated as an image.

Our main contributions in this paper are:

- Determining the signal keypoints to be transmitted to obtain an accurate TDOA estimation, at lower data rates or improved accuracy as against GCC solutions.
- Demonstrating the robustness of the proposed technique to different noise and reverberation conditions.
- Comparison of the proposed technique with another data-driven approach, namely audio fingerprinting.

II. Methodology

The proposed approach is based on Fig. 1, in which keypoint extraction occurs at the sensor-head. These keypoints are then communicated to a fusion center, which may either be a centralised node, or simply another sensor node. The communications channel is assumed to be low-bandwidth, such that minimal communication is desirable to ensure low-latency in the full localisation system. The sensors considered in this paper are microphones, but could naturally be any passive transducer, such as hydrophones, or RF.

The sensors $\mathbf{s}_i$ and $\mathbf{s}_j$ measure signals, m_i and m_j. The proposed algorithm for estimating TDOA, for that pair of microphones, consists of the following key steps:

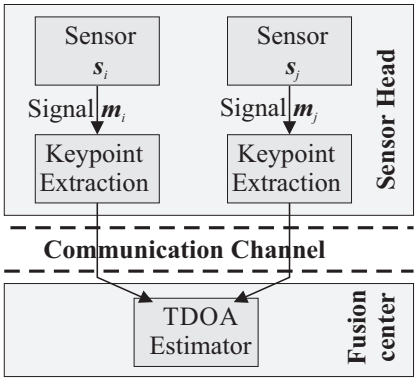


Fig. 1. Overview of the system architecture.

1) **At the Sensor-Head:** Calculate the spectrograms, $\widetilde{m}_i$ and $\widetilde{m}_j$ at each microphone, from the recorded signals m_i and m_j. The dimension of each spectrogram is F by T, where F is the number of rows corresponding to frequencies and T is the number of columns corresponding to time. We determined the optimum parameters for calculating the spectrogram were window size = 256, overlap = 204 and the final number of sampling points in the discrete Fourier transform = 1024;
2) **At the Sensor-Head:** Calculation of the Scale-Invariant Feature Transform (SIFT) [19] on the normalized spectrogram magnitude, in order to detect n keypoints from each spectrogram. We create a vector of keypoints, $\mathbf{f_i}$ and $\mathbf{t_i}$ for the *i-th* microphone. The k^{th} keypoint has coordinates (f_k, t_k), which corresponds to the time-frequency location at which the keypoints are detected. The values that will be transmitted are integers (encoded in 32 bits in order to keep high precision) and we only need to transmit the $t-$coordinates. It was found that adding in the frequency information did not improve the Time-Difference of Arrival (TDOA) relative error. Therefore, the total number of data samples that need to be transmitted to the fusion center is $n \times 32$. We experimented with the number of keypoints that need to be transmitted in order to obtain an acceptable margin of error in the Time-Difference of Arrival (TDOA). In light of this, we selected keypoints with the highest energy frequency coefficients, i.e. points that belong to rows of the spectrogram in which the sum of coefficients at key points is large. We selected k-rows each time, where k varies between 0.1 and 1;
3) **At the Fusion Center:** After the data is transmitted, two new vectors, $\widehat{\mathbf{m_i}}$ and $\widehat{\mathbf{m_j}}$, of the same size as m_i and m_j are created at the fusion center. We are assuming that all the sensors are synchronised and therefore started recording at the same instant. We can map keypoint locations to vectors by pre-calculating the times that correspond to the t-coordinates. The vector is filled with 1's in indices where a SIFT keypoint was detected and with 0's otherwise;

$$\widehat{\mathbf{m_i}}(l) = \begin{cases} 1 & \text{if } l \in \mathbf{t_i} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

4) **At the Fusion Center:** Calculation of Generalised Cross-Correlation (GCC) (defined by the $\star$ operator) between both vectors in the time domain. Since the cross-correlation is now on a binary vector, there is no need for the spectral normalisation as in PHAT.

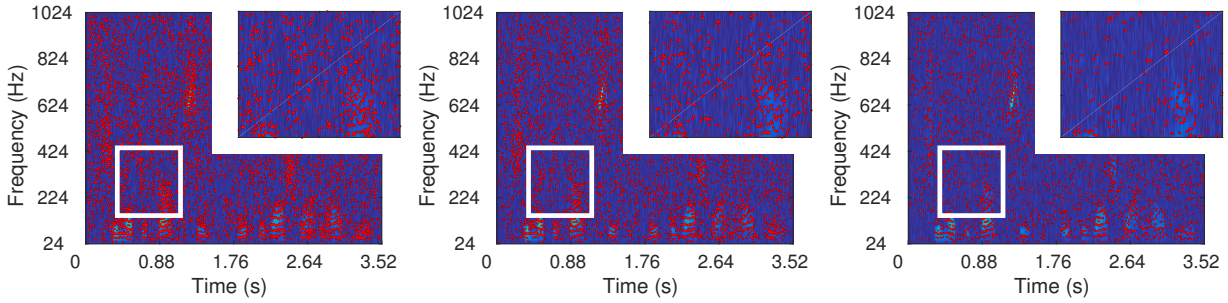
$$\tau_{\text{delay}} = \arg\max_t((\widehat{\mathbf{m_i}} \star \widehat{\mathbf{m_j}})(t)) \tag{2}$$

III. Experimental Results

We performed experiments using speech signals from the TIMIT database [20] and simulated environments by means of the image-source method [21]. We simulated two microphones in a linear array, separated by a distance of 4 metres and sampled at 16kHz. The simulated room has a size of 25m $\times$ 3m $\times$ 12m.

Since Time-Difference of Arrival (TDOA) is in the order of milliseconds for some source locations and centiseconds for others, it is necessary to standardize the error in order to make a fair comparison among source positions. Using the Ground Truth (GT), the relative error is computed using the TDOA estimation error in Equation 3. Similarly, we use the same principle to estimate the Direction of Arrival (DOA) relative error in Eq. 4.

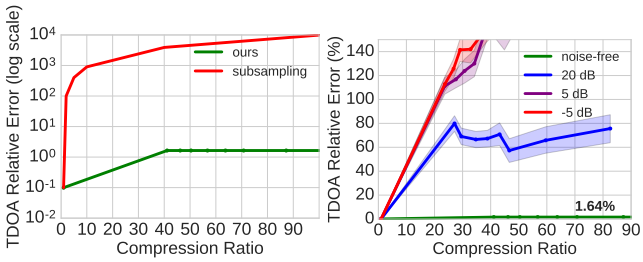
$$\text{tdoa error}(\%) = \frac{\|\text{tdoa} - \text{gt}\|}{\|\text{gt}\|} * 100 \tag{3}$$



(a) $40:1$ compression (b) $55:1$ compression (c) $90{:}1$ compression

Fig. 2. SIFT keypoints (indicated in red) in the signal spectrogram, for different compression ratios. For each spectrogram, a patch (white rectangle) is selected and magnified at the upper right corner to provide a clearer visualization of the SIFT keypoints. This illustrates how the selected SIFT features are not necessarily spectrogram peaks and how our features differ from the peak picker approaches.

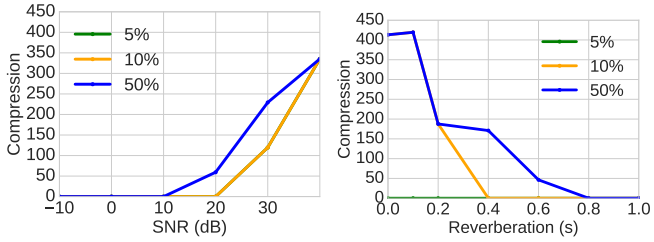
As previously mentioned, the compression ratio was varied in order to determine how much compression we can achieve while obtaining a reasonable TDOA relative error. We used the subsampling strategy presented in Sec. II, where we selected keypoints with the highest energy frequency coefficients, i.e points that belong to rows of the spectrogram in which the sum of coefficients at key points is large. Fig. 2 shows the spectrogram SIFT keypoints for different compression ratios. Fig. 3 illustrates the TDOA relative error with respect to compression ratio. In this experiment, the source was located at a DOA of 45°. Fig. 3a shows the error for an environment free of noise and reverberation using the proposed method and compares it with an approach in which compression is achieved by subsampling the signal. Since subsampling the signal increases the error dramatically even for low compression ratios, we decided to use a logarithmic scale on the Y-axis. Fig. 3b shows the relative error for a non-reverberant environment for different levels of noise. For a signal with SNR 20dB the TDOA error remains below 100%.



(a) Subsampling vs Ours (b) Various Noise Levels

Fig. 3. TDOA Relative Error achieved for different compression ratios for a source located at DOA 45°. The figure of the left shows the TDOA relative error for our algorithm compared with a baseline in which the signal is compressed by subsampling. We used the logarithmic scale on the Y-axis given that the error for the subsampling approach is much higher than our error. The right-hand side of the figure shows the TDOA Relative Error for a noise-free signal and for signals with various SNR values. To estimate the relative error for each compression ratio, we used 100 simulations.

Fig. 4 shows how noise and reverberation separately affect the compression ratio. We calculated the minimum value of compression that produced a TDOA relative error smaller

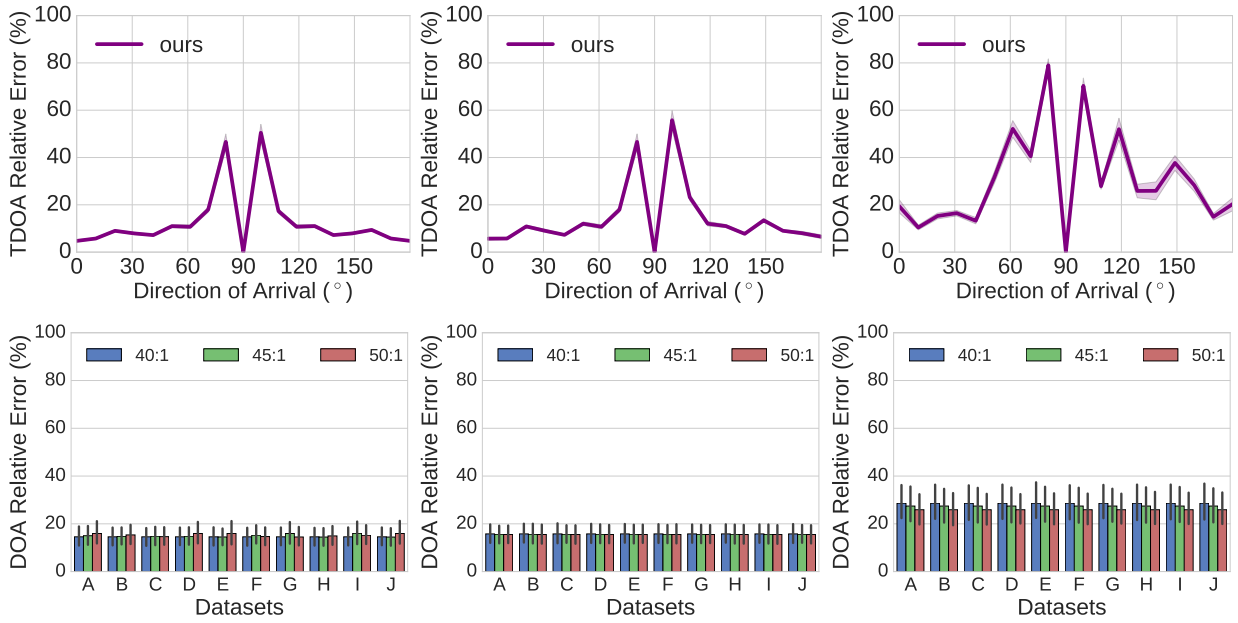


(a) Noise (b) Reverberation

Fig. 4. Maximum compression when the TDOA relative error $\leq 5\%, 10\%, 50\%$ for a source located at DOA $45°$ for different values of noise and reverberation. In 4(a), white Gaussian noise of -10 dB, 0 dB, 10 dB, 20 dB, 30 dB and 40 dB signal-to-noise ratio per sample was added to the original signal. For 5% and 10%, the compression ratios are identical, therefore we can only visualize a single line. In 4(b), we simulated reverberation values of $T_{60} = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$ seconds.

than $5\%, 10\%, 50\%$ for the given noise and reverberation conditions. In this scenario, the source is located at DOA $45°$. In Fig. 4(a), a white Gaussian noise of -10 dB, 0 dB, 10 dB, 20 dB, 30 dB and 40 dB signal-to-noise ratio per sample was added to the original signal. Note how the compression improves as the signal-to-noise ratio (SNR) gets higher. In the case, of 5% and 10%, the compression ratios are identical, therefore we can only visualize one line. We used T_{60} as a measurement of reverberation, interpreted as the time it takes a signal to drop by 60dB. In Fig. 4(b), reverberation values of $T_{60} = \{0.1k,\ k \in \{1, \ldots, 10\}\}$ seconds are simulated. In this case we can see that there is no compression value for which the error is smaller than 5%, however for 10% and 50% we achieved high compression ratios for low reverberation values (up to 0.6), after which the compression decreases to zero.

Fig. 5 illustrates the TDOA relative error and the DOA relative error for 3 different reverberation levels: $T_{60} = \{0.1, 0.2, 0.3\}$ seconds. We randomly selected 10 different sounds from the TIMIT dataset, which included speech signals from 5 men and 5 women (labeled A to J). We simulated 19 different source locations (DOA), from 0° to 180°, with a step size of 10°. We ran 5 different simulations for each of these sources and reverberation values. The first row of



(a) reverberation $T_{60} = 0.1$ (b) reverberation $T_{60} = 0.2$ (c) reverberation $T_{60} = 0.3$

Fig. 5. TDOA relative error and the DOA relative error for 3 different reverberation levels: $T_{60} = \{0.1, 0.2, 0.3\}$ seconds. The results are from 10 speech signals (labelled A to J), at 19 different locations (DOA), from 0°to 180°, with a step size of 10°. We ran 5 different simulations for each of these sources and reverberation values. The first row shows the TDOA relative error for each DOA. The compression ratio is $40:1$ for each signal. The second row shows the DOA localization error per dataset for three different compression ratios: $40:1$, $45:1$ and $50:1$.

Fig. 5, shows the TDOA relative error for each DOA. The compression ratio is $40:1$ for each signal. It can be seen from the plots that for environments with low reverberation, $T_{60} = 0.1, 0.2$ seconds, the TDOA relative error is smaller than 20% for most DOA, except for 80°and 100°, in which case the error rises above 40%. The reason for this behavior is the small values of TDOA at such locations, which makes its calculation very challenging. The second row of Fig. 5 shows the DOA localization error. The x-axis presents 10 different datasets (labeled A to J). Three different compression ratios are used: $40:1$, 45:1 and 50:1. For low reverberation, $T_{60} = 0.1, 0.2$ seconds, the DOA relative error remains less than 20% for different compression ratios and sources. When reverberation $T_{60} = 0.3$ seconds, the TDOA relative error increases dramatically for most DOA, especially for 80°and 100°, in which case it is close to 80%. This large TDOA error has little impact on the DOA estimation, however. Even though the DOA relative error is above 20% in this case, the error in general remains less than 40%.

$$\text{doa error}(\%) = \frac{\|\text{doa} - \text{gt}\|}{\|\text{gt}\|} * 100 \tag{4}$$

IV. Discussion

We found that, by applying computer vision techniques, Scale-Invariant Feature Transform (SIFT), to the spectrogram of a speech signal, it is possible to detect keypoints that contain relevant information about the signal. We were able to use these keypoints to select the signal samples used to estimate Time-Difference of Arrival (TDOA) within a reasonable margin of relative error.

Our mechanism for improving the compression rate is to use subsampling of the SIFT keypoints in the spectrogram constructed at each sensor (microphone). Our strategy was to select the highest energy frequency coefficients, i.e rows of the spectogram in which the sum of coefficients at key points is large. This proved to be effective in scenarios in which there is little noise, as illustrated by Fig. 3b and Fig. 4a.

We ran our algorithm for various source locations and speech signals. We determined that the highest error in estimating the TDOA was caused in positions where the source was located in front of the microphone array, either at 80°or 100°. This happens because the TDOA is very small for these positions, which complicates the estimation. For 90°, where the TDOA is zero, and for 0°and 180°, where the separation is maximum, the relative error is closer to zero. On the other hand, given a similar position and the same noise and reverberation conditions, our algorithm performs very similarly across the test speech signals we used.

The algorithm's main drawback is its sensitivity to reverberation, as is evidenced in Fig. 4 and Fig. 5. This may be attributable to SIFT keypoints chosen from reverberations rather than from the original signal. One strategy to overcome this problem might be to estimate the probability of the keypoints being reverberations based on the amplitude of neighboring keypoints.

Table I shows a comparison between our algorithm and audio fingerprinting [15] for TDOA estimation. Both algo-

TABLE I
FINGERPRINT VS OURS: TDOA RELATIVE ERROR

	Fingerprint Noise (SNR)			Ours Noise (SNR)		
Reverb	**noise-free**	**-5 dB**	**-10 dB**	**noise-free**	**-5 dB**	**-10 dB**
0	75.93 (43.80)	140.86 (71.35)	115.38 (55.48)	0.94 (0.68)	114.94 (38.62)	102.77 (12.05)
0.1	77.61 (50.29)	136.34 (69.54)	108.78 (59.71)	0.94 (0.68)	94.92 (29.44)	97.84 (20.87)
0.2	80.10 (40.72)	143.88 (69.69)	115.22 (60.60)	1.33 (0.69)	97.84 (14.91)	99.08 (8.68)
0.4	86.90 (43.88)	147.81 (80.15)	121.74 (63.04)	18.81 (35.14)	107.08 (13.68)	94.61 (12.13)
0.6	89.82 (92.69)	149.26 (76.36)	129.49 (65.35)	36.13 (26.12)	87.06 (31.63)	99.38 (20.24)

rithms were run on 10 different speech signals from the TIMIT database [20]. A value of noise (from 0 to -10 dB) and reverberation (from 0 to 0.6) was added to the signal. For each of these noise and reverberation values, the algorithm was executed 50 times. The table presents the mean TDOA relative error with the standard deviation (in brackets). In this scenario, the source was located at DOA 45°. We used the implementation of audio fingerprinting presented in [21], in which the input signal is subsampled to 8kHz to calculate the spectrogram. The number of sections is 64ms and the overlap is 32ms. We selected 50 landmarks per signal to perform our comparison. Table I shows how audio fingerprinting error is larger than ours for this particular source location and these particular speech signals.

V. CONCLUSIONS AND FUTURE WORK

In this work, we showed that, by applying a computer vision approach to the spectrogram of a speech signal, it was possible to identify samples of the signal allowing for an estimation of Time-Difference of Arrival (TDOA) within a reasonable margin of relative error. We tested the robustness of the proposed technique under different noise and reverberation conditions using different speech signals and source locations. We showed that our algorithm can estimate TDOA and the source location within an acceptable error range when the compression ratio of the signal is $40:1$.

In the future, we plan to modify our algorithm by improving on its robustness to noise and reverberation. We intend to do this by estimating the probability of keypoints representing reverberation or not depending on the amplitude of its neighbors. Moreover, we would like to perform experiments in open spaces in order to evaluate how the high reverberation values affect our algorithm.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[2] H.-s. Wang, J. Li, Z.-q. Sun, M.-h. Cao, and H.-w. Xie, "Accurate delay extraction for indoor pulse sound source location," in *Signal Processing (ICSP), 2014 12th International Conference on*. IEEE, 2014, pp. 298–301.

[3] P. Pertila, M. S. Hamalainen, and M. Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 11, pp. 2393–2402, 2013.

[4] H.-K. Hao, H.-M. Liang, and Y.-W. Liu, "Particle methods for real-time sound source localization based on the multiple signal classification algorithm," in *Intelligent Green Building and Smart Grid (IGBSG), 2014 International Conference on*. IEEE, 2014, pp. 1–5.

[5] Y. G. Kim, K. M. Jeon, Y. Kim, C.-H. Choi, H. K. Kim, and L. Nex, "Underwater acoustic sensor array signal lossless compression based on valid channel decision approach," *Int J Image Signal Syst Eng*, vol. 1, no. 1, pp. 21–28, 2017.

[6] S. Zhou and L. Ying, "On delay constrained multicast capacity of large-scale mobile ad hoc networks," *IEEE Transactions on Information Theory*, vol. 61, no. 10, pp. 5643–5655, 2015.

[7] G. Simon and L. Sujbert, "Acoustic source localization in sensor networks with low communication bandwidth," in *Intelligent Solutions in Embedded Systems, 2006 International Workshop on*. IEEE, 2006, pp. 1–9.

[8] Q. Fuyong, G. Fucheng, J. Wenli, and M. Xiangwei, "Data compression based on DFT for passive location in sensor networks," *Procedia Engineering*, vol. 29, pp. 3091–3095, 2012.

[9] D. O. Zion and H. Messer, "Envelope only tdoa estimation for sensor network self calibration," in *Sensor Array and Multichannel Signal Processing Workshop (SAM), 2014 IEEE 8th*. IEEE, 2014, pp. 229–232.

[10] N. El Gemayel, H. Jakel, and F. K. Jondral, "Error analysis of a low cost tdoa sensor network," in *Position, Location and Navigation Symposium-PLANS 2014, 2014 IEEE/ION*. IEEE, 2014, pp. 1040–1045.

[11] R. Sonnleitner and G. Widmer, "Robust quad-based audio fingerprinting," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 3, pp. 409–421, 2016.

[12] S. Baluja and M. Covell, "Waveprint: Efficient wavelet-based audio fingerprinting," *Pattern recognition*, vol. 41, no. 11, pp. 3467–3480, 2008.

[13] T. Tsai and A. Stolcke, "Robust and efficient multiple alignment of unsynchronized meeting recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 833–845, 2016.

[14] A. Wang *et al.*, "An industrial strength audio search algorithm." in *Ismir*, vol. 2003. Washington, DC, 2003, pp. 7–13.

[15] T.-K. Hon, L. Wang, J. D. Reiss, and A. Cavallaro, "Audio fingerprinting for multi-device self-localization," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 10, pp. 1623–1636, 2015.

[16] M. Zanoni, S. Lusardi, P. Bestagini, A. Canclini, A. Sarti, and S. Tubaro, "Efficient music identification approach based on local spectrogram image descriptors," in *Audio Engineering Society Convention 142*. Audio Engineering Society, 2017.

[17] X. Zhang, B. Zhu, L. Li, W. Li, X. Li, W. Wang, P. Lu, and W. Zhang, "Sift-based local spectrogram image descriptor: a novel feature for robust music identification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 6, 2015.

[18] Y. Ke, D. Hoiem, and R. Sukthankar, "Computer vision for music identification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 597–604.

[19] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus, 1993," *Linguistic Data Consortium, Philadelphia*.

[21] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.