On the use of acoustic flow sensors for far-field audio capture

Frederic Lepoutre, Lucas Henrique Teixeira Carneiro, Philippe-Aubert Gauthier

ON THE USE OF ACOUSTIC FLOW SENSORS FOR FAR-FIELD AUDIO CAPTURE

Frédéric Lepoutre[1] **Lucas Carneiro**[2] **Philippe-Aubert Gauthier**[3]

[1] Soundskrit Inc., Montréal, QC, Canada

[2] Faculté de Génie, Université de Sherbrooke (GAUS), Sherbrooke, QC, Canada

[3] Ecole des arts visuels et médiatiques, Université du Québec, Montréal, QC, Canada

frederic.lepoutre@soundskrit.ca

lucas.carneiro@usherbrooke.ca

gauthier.philippe-aubert@uquam.ca

ABSTRACT

Omni-directional microphone arrays have enabled the robust capture of speech in noisy and reverberant environments. However, for good performance, it requires many microphones over a wide aperture which is impractical for many applications. A different approach consists in using an array of co-located directional microphones. While rarely implemented in mass market due to hardware constraints, these systems do not suffer from the physical limitations of aperture effects. Recently, a new bio-inspired approach to acoustic velocity sensing was proposed and shows promising features for the design of small, high performance directional microphones. In this paper, an overview of microphone arrays and their limitations, as compared to directional sensors solutions that could be built with the new acoustic flow sensor, is exposed.

1. INTRODUCTION

In recent years, new applications of voice capture emerged in the consumer electronics industry. Whether it is a smart speaker being able to understand voice command of a user, or a smartphone being able to function as a conference phone from a table, smart devices are now capable of capturing speech signals in noisy and reverberant environments. What enabled these functionalities is the widespread adoption of MEMS microphones arrays. Indeed, by combining multiple microphone signals together, a virtual directional microphone can be steered towards a sound source of interest and reject sounds coming from other directions. This allows a clean speech signal capture, with reduced reverberation and reduced background noise. This cleaner signal enables highly intelligible conversations content in phone calls and voice recognition with higher accuracy.

However, the performance of a typical microphone array is limited by the aperture effect. On the other hand, arrays made of co-located directional capsules such as ambisonics microphones or acoustic vector sensors do not suffer from these limitations. As of today, these co-located systems are not ubiquitous as the required microphone capsules are big or impractical [1]. But recently, a new concept for acoustic velocity sensing have been presented inspired by the auditory systems of small insects [2]. This sensor can be used to build multi-directional microphones with differentiating properties while suitable for mass markets. It could address associated far-field audio capture applications.

The first section of this paper consists of an overview of the basic concepts of microphone arrays as well as practical limitations. Directional microphone arrays are then presented in the second section. Its advantages compared to microphone arrays are explained.

2. OVERVIEW OF MICROPHONE ARRAYS

A microphone array is made of omni-directional capsules ("omnis") placed at different position. Depending on where the sound sources are located, the emitted sound waves hit the microphones with a slight time difference. The purpose of array processing is to leverage these time differences to detect, reject or enhance sounds coming from specific directions.

In the following, an acoustic model using a 3-microphone array is presented.

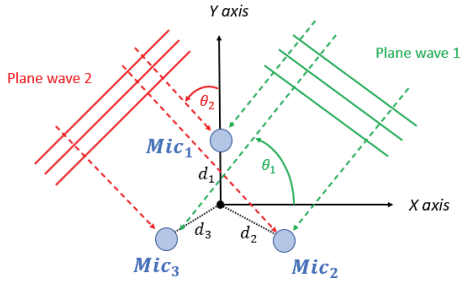


Figure 1. Illustration of a circular 3-microphone array with two incoming planar waves.

Assuming there are two incoming plane waves coming from angle θ_1 and angle θ_2, each microphone outputs a signal $mic_i(t)$ that captures a mixture of acoustic signals $s_i(t)$ and microphone electrical noise signal $n_i(t)$. Noise signals are uncorrelated with the sound waves and with each other:

$$\begin{cases} mic_1(t) = s_1(t-\tau_{1,1}) + s_2(t-\tau_{1,2}) + n_1(t), \\ mic_2(t) = s_1(t-\tau_{2,1}) + s_2(t-\tau_{2,2}) + n_2(t), \\ mic_3(t) = s_1(t-\tau_{3,1}) + s_2(t-\tau_{3,2}) + n_3(t). \end{cases} \quad (1)$$

When hit by a sound wave i, each microphone j outputs delayed versions of the signal s_i by a time $\tau_{i,j}$ due to the slight distance difference that the sound wave has to travel before it reaches each microphone. In the frequency domain, Equation (1) gives:

$$\begin{cases} MIC_1(\omega) = S_1(\omega)e^{-j\omega\tau_{1,1}} + S_2(\omega)e^{-j\omega\tau_{1,2}} + N_1(\omega), \\ MIC_2(\omega) = S_1(\omega)e^{-j\omega\tau_{2,1}} + S_2(\omega)e^{-j\omega\tau_{2,2}} + N_2(\omega), \\ MIC_3(\omega) = S_1(\omega)e^{-j\omega\tau_{3,1}} + S_2(\omega)e^{-j\omega\tau_{3,2}} + N_3(\omega). \end{cases} \quad (2)$$

2.1 Beamforming

Beamforming consists in combining microphone outputs to recover the desired signal s_1 while rejecting the interferer s_2 and all noise sources:

$$\begin{aligned} b(t) &= w_1.mic_1(t) + w_2.mic_2(t) + w_3.mic_3(t) \\ &= s_{1,out}(t) + s_{2,out}(t) + n_{out}(t). \end{aligned} \quad (3)$$

The resulting output contains acoustic signals s_1 and a residual of s_2 and noise signals. The performance of the beamformer can be measured by its output signal to total noise ratio (STNR), which is the energy of the desired signal in the output over the energy of all other signals. Similarly, it is defined the output signal to noise ratio (SNR) and output signal to interferer ratio (SIR), as in Benesty [3]:

$$\begin{cases} SIR_{out} = \frac{E\left[|s_{1,out}|^2\right]}{E\left[|s_{2,out}|^2\right]}, \\ SNR_{out} = \frac{E\left[|s_{1,out}|^2\right]}{E[|n_{out}|^2]}, \\ STNR_{out} = \frac{E\left[|s_{1,out}|^2\right]}{E\left[|s_{2,out}|^2\right]+E[|n_{out}|^2]} = \frac{SIR_{out}\,.\,SNR_{out}}{SIR_{out}+SNR_{out}}, \end{cases} \quad (4)$$

where E is the mathematical expectation. The same metrics are defined for the raw output of the microphone:

$$\begin{cases} SIR_{mic} = \frac{E[|s_1|^2]}{E[|s_2|^2]}, \\ SNR_{mic} = \frac{E[|s_1|^2]}{E[|n_1|^2]}, \\ STNR_{mic} = \frac{E[|s_1|^2]}{E[|s_2|^2]+E[|n_1|^2]} = \frac{SIR_{mic}\,.\,SNR_{mic}}{SIR_{mic}+SNR_{mic}}. \end{cases} \quad (5)$$

By mixing the microphones with the weights w_i, a virtual, directional microphone is created. The choice of w_i controls the shape of the polar pattern, the look direction, and its sensitivity. There are different strategies to design the weights w_i: the polar pattern may be made as narrow as possible to reject s_2 (maximizing SIR_{out}), or to capture s_1 with maximum sensitivity by rejecting microphone self-noise n (maximizing SNR_{out}). Unfortunately, these two goals are antagonists [3] and one must trade-off between maximizing SIR_{out} and SNR_{out} to maximize $STNR_{out}$.

To make the array as directional as possible (i.e. maximize SIR_{out}), one way is to follow the differential microphone array (DMA) approach [3]. The technique consists in pairing microphones together and subtracting the two outputs to cancel sounds coming from given directions. In Fig. 2, this process is illustrated using two microphones mic_1 and mic_2. To reject sounds coming orthogonally to the axis made by the two microphones, it is intuitive to subtract them, as the sound wave reaches the two microphones at the same time, leading to complete cancellation. Sounds coming from other directions hit the pair of microphones with different timings, therefore they are not completely cancelled by the subtraction[1]. The associated polar pattern is a first-order dipole with a null toward the 30-degree direction.

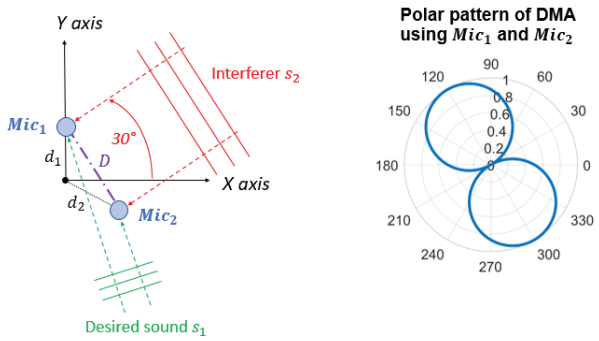
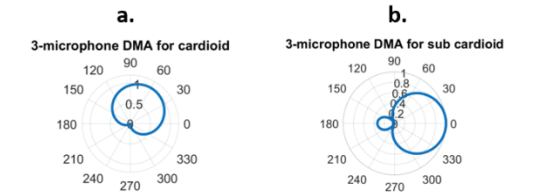


Figure 2. On the left, illustration of a 2-microphone array where a DMA is used to cancel an interferer coming at 30 degrees. On the right, the obtained directivity.

Assuming ideal plane waves, no matter how small of an angle there is between the sound source s_1 relative to s_2, s_2 is always cancelled, and s_1 is always captured: the DMA microphone is infinitely directional for this interferer, which translates into infinite SIR_{out}. To rotate the polar pattern, a time delay can be added to one of the microphones to ensure that the signals coming from any desired null direction are perfectly in phase before the subtraction stage. When many microphones are involved, the DMA design methodology proposed by Benesty [3] can be followed to create different polar patterns, with various directivity and look direction, as illustrated in Fig. 3.



Coefficients for cardioid	
w_1	0.16698947 + 1.83697337i
w_2	0.16698947 + 1.83697337i
w_3	0.16712418 - 3.68449105i

Coefficients for sub-cardioid	
w_1	0.00016347 + 4.73811756i
w_2	0.16695255 - 2.36906454i
w_3	0.16695255 - 2.36906454i

Figure 3. Choice of weights using the DMA methodologies at 125 Hz to design a cardioid (a.) and a sub-cardioid (b.).

The DMA method creates highly directional beams. However, directionality comes at a cost of a highly

[1] This is only true up to a certain frequency due to spatial aliasing, as will be explained later.

attenuated speech signal while the self-noise of the sensors has been amplified, leading to a decreased overall SNR_{out}. This is illustrated in the following for the situation of Fig. 2. The equations are:

$$\begin{cases} mic_1(t) = s_1(t-\tau_{1,1}) + s_2(t-\tau) + n_1(t), \\ mic_2(t) = s_1(t-\tau_{2,1}) + s_2(t-\tau) + n_2(t). \end{cases} \quad (6)$$

The sound emitted by s_2 reaches the two microphones mic_i at the same time so the two time delays $\tau_{i,2}$ are equal to the same value, defined as τ. Consequently, the output of the dipole DMA in the time domain and in the frequency domain is expressed below:

$$\begin{aligned} b(t) &= mic_1(t) - mic_2(t) \\ &= s_1(t-\tau_{1,1}) + s_1(t-\tau_{2,1}) + n_1(t) - n_2(t), \end{aligned} \quad (7)$$

$$\begin{aligned} B(\omega) &= S_1(\omega)\left[e^{-j\omega\tau_{1,1}} - e^{-j\omega\tau_{2,1}}\right] + N_1(\omega) - N_2(\omega), \\ B(\omega) &= S_1(\omega)e^{\frac{-j\omega(\tau_{1,1}+\tau_{2,1})}{2}}\left[2j.\sin\left(\frac{\omega(\tau_{2,1}-\tau_{1,1})}{2}\right)\right] + \\ &\quad N_1(\omega) - N_2(\omega). \end{aligned} \quad (8)$$

leading to the following energy of the beamformer:

$$\begin{aligned} E[|B(\omega)|^2] &= 4.E[|S_1(\omega)|^2].\sin\left(\frac{\omega(\tau_{2,1}-\tau_{1,1})}{2}\right)^2 + \\ &\quad E[|N_1(\omega)|^2] - E[|N_2(\omega)|^2]. \end{aligned} \quad (9)$$

Linking the time difference $\tau_{2,1} - \tau_{1,1}$ to the spacing D between microphones, the angle of arrival θ_1 and the speed of sound c, and assuming the energy of the noise of all microphone is the same and equal to $E[|N(\omega)|^2]$, it is possible to compute SNR_{out} and express it as a gain applied to SNR_{in}:

$$SNR_{out} = 2.\sin\left(\frac{\omega}{2}\left(\frac{D}{c}\sin(\theta_1 - 30°)\right)\right)^2 . \frac{E[|S_1(\omega)|^2]}{E[|N(\omega)|^2]}, \quad (10)$$

$$\frac{SNR_{out}}{SNR_{mic}} = 2.\sin\left(\frac{\omega}{2}\left(\frac{D}{c}\sin(\theta_1 - 30°)\right)\right)^2 . \quad (11)$$

Hence, the relationship between SNR_{in} and SNR_{out} is highly dependent on the microphone spacing and the frequency. At large wavelength compared to the spacing, the sensitivity drops by 40 dB per decade, as illustrated in Fig. 4. At higher wavelength, there is comb-filtering as will be explained later. In between, there are transition frequencies where sensitivity is boosted up to 6 dB.

While the background noise has been completely rejected, leading to an infinite SIR_{out}, SNR_{out} is mostly lower than SNR_{in}. Overall, the output STNR is lower than the input STNR at most frequencies. This situation is displayed in Fig. 5 assuming an input SNR of 4 dB and SIR_{in} of 1 dB.

The highly directional DMA is only beneficial for a limited frequency range, as it boosts the self-noise of the sensors. For other frequencies, a different beamforming strategy is required, one that will aim at improving the SNR with the trade-off less SIR improvement.

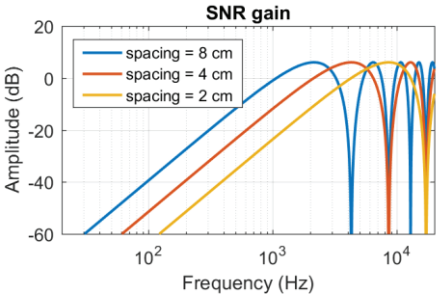


Figure 4. Display of the SNR gain as a function of frequency and microphone spacing in the case of a dipole DMA using two microphones.

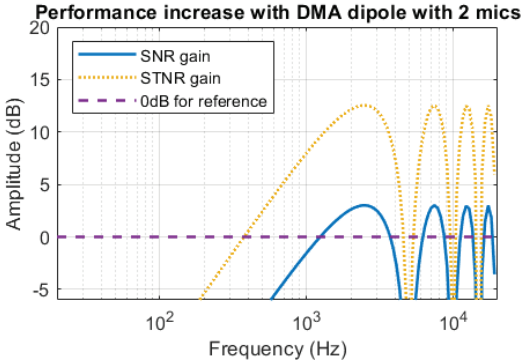


Figure 5. Display of performance gain as a function of frequency using a two-microphone DMA with 4 cm spacing. The SIR gain is not displayed as it is infinite.

The beamformer with the highest output SNR is called the delay-and-sum beamformer (DASB). The idea is to apply a delay to each microphone to time align sounds coming from a specific direction. Then all microphones are summed together, leading to constructive interference of the desired sound. Sounds coming from other directions will be somewhat attenuated, as their waveforms are not time aligned before summation. As the self-noise signals are not time aligned, they are not boosted as much as the desired sound source. The equation for the beamformer that maximizes the sensitivity for the signal s_1 is:

$$\begin{aligned} b(t) &= mic_1(t-\tau_{2,1}-\tau_{3,1}) + mic_2(t-\tau_{1,1}-\tau_{3,1}) + \\ &\quad mic_3(t-\tau_{1,1}-\tau_{2,1}), \\ b(t) &= s_{1,out}(t) + s_{2,out}(t) + n_{out}(t). \end{aligned} \quad (12)$$

In the frequency domain, this gives:

$$B(\omega) = S_{1,out}(\omega) + S_{2,out}(\omega) + N_{out}(\omega), \quad (13)$$

where the terms are:

$$\begin{aligned} S_{1,out}(\omega) &= 3.S_1(\omega)e^{-j\omega(\tau_{1,1}+\tau_{2,1}+\tau_{3,1})}, \\ S_{2,out}(\omega) &= S_2(\omega)\left[e^{-j\omega(\tau_{2,1}+\tau_{3,1})} + e^{-j\omega(\tau_{1,1}+\tau_{3,1})} + \right. \\ &\quad \left. e^{-j\omega(\tau_{1,1}+\tau_{2,1})}\right], \\ N_{out}(\omega) &= N_1(\omega)e^{-j\omega(\tau_{2,1}+\tau_{3,1})} + N_2(\omega)e^{-j\omega(\tau_{1,1}+\tau_{3,1})} + \\ &\quad N_3(\omega)e^{-j\omega(\tau_{1,1}+\tau_{2,1})}. \end{aligned} \quad (14)$$

The beamformer will increase the SNR and SIR. However, the polar pattern is not as directional as obtained by the DMA method. Additionally, the polar pattern is highly wavelength dependent, as the DASB is barely directional at low wavelength and progressively becomes directional at higher frequencies, as shown in Fig. 6.

As is shown in Fig. 7, the DASB offers an increase in SNR across all frequencies, and an increase in SIR at higher frequencies. The output SIR is lower than the DMA, but the output SNR is higher.

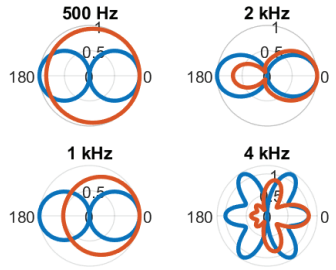


Figure 6. Comparison of polar patterns of a dipole DMA (blue) and DASB (red), using 3 microphones with 4 cm spacing.

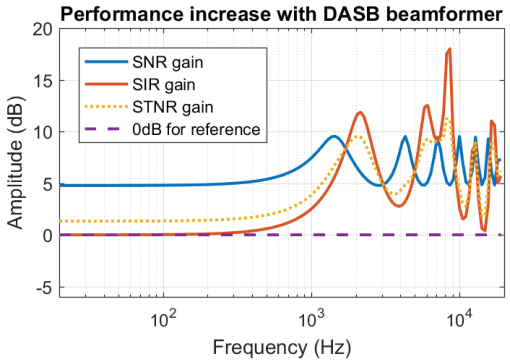


Figure 7. Comparison of SNR, SIR and STNR gains for the DASB, using 3 microphones with 4 cm spacing.

Therefore, one should choose different beamforming strategies depending on the level of sensor noise compared to the desired signal (SNR oriented strategy) and the level of interferer (SIR oriented strategy). The DMA is more directional and therefore better suited to reject interferers. The DASB maximizes the output SNR and is better suited to reject the microphone noise and therefore capture quiet distant sound. Between these two extreme cases, design methodologies can trade-off between directionality and sensitivity [3].

Beamforming only works properly when there is a one-to-one correspondence between time difference of arrivals at microphones and a given direction in space. This is only true at wavelength larger than the microphone array. At higher frequencies, an ambiguity appears, as a single time difference of arrival (TDoA) between microphones can correspond to multiple directions [4]. This translates into aliasing side lobes in the directivity pattern, no matter what beamforming method is used, as can be seen on Fig. 6 at 8 kHz and on Fig. 4, 5 and 7 at higher frequencies.

Because of the side lobes, microphone arrays typically do not operate above aliasing frequency, which puts constraints on the maximum spacing between microphones. At the same time, the spacing must be kept large enough to mitigate the sensor noise amplification. Overall, a good trade-off is obtained when spacing is around 4 cm to operate on the most important speech frequencies (between 300 Hz and 4 kHz).

2.2 Localization

As the beamformer needs to be aimed properly to capture or reject sound from a given angle of arrival (AoA), it is critical to localize the sound of interests. There exists three categories of localization methods: those based on triangulation using the estimated time-differences of arrival between pairs of microphones, those using a beamformer to search for the AoA that maximizes the output power, and those leveraging the geometrical relationships between the sub-spaces spanned by the eigen vectors of the signal covariance matrix, such as the MUSIC algorithm [5]. In the following is explained the time-delay estimate method for a single pair of microphones.

Taking the model of planar waves in Fig. 1, the TDoA δ_t between the microphone 2 and 3 associated with the plane wave 1 is linked to its angle of arrival as:

$$\theta_1 = \text{acos}\left(\frac{\delta_t . c}{D}\right). \tag{15}$$

Therefore, the localization problem boils down to the accurate estimation of the TDoA. For this task, the most popular method is to look for the time argument τ that maximizes the cross-correlation function as defined by:

$$R_{mic_2,mic_3}(\tau) = E[mic_2(t).mic_3(t-\tau)], \tag{16}$$

where E is the expectation. Indeed, the cross-correlation can be shown to be a Dirac function at abscissa dt spread by the autocorrelation of the plane wave signal $s_1(t)$. This method is presented as the generalized cross-correlation (GCC) by Knapp [6]. Therefore, δ_t can be found as the abscissa of the peak, as illustrated in Fig. 8.

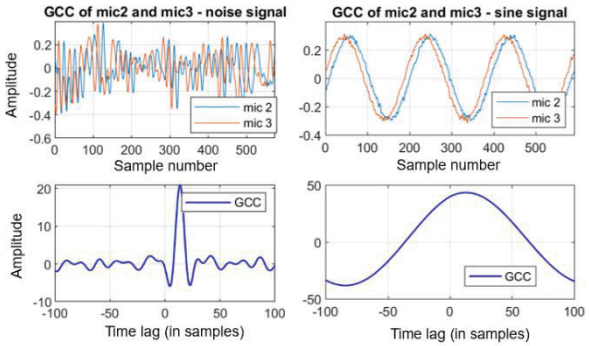


Figure 8. Comparison of cross-correlation curves when $s_1(t)$ is white noise (left) and a sine wave at 500 Hz (right), for spacing of 4 cm. Sensor noise is added for illustration with SNR of 20 dB.

In Fig. 8, it can be seen that the peak is sharper for white noise, but both peaks occur at a time lag value corresponding to the time delay between the waveforms (here expressed in samples at a sampling rate of 16kHz, multiplying it by the sampling rate would yield the time in seconds). The peak of the correlation must be estimated with high time accuracy, as a slight offset can impact the AoA estimation dramatically [4]. Therefore, interpolation and upsampling on the waveforms is typically applied beforehand. Additionally, spatial aliasing prevents localization at high frequencies [7].

2.3 Direct-diffuse sound separation

Microphone arrays have also the ability to discriminate between direct and diffuse sounds. The acoustic signals are made of a mixture of direct sounds coming from specific directions, and unwanted background noise randomly coming from all directions such as reverberation. As reverberation decreases the speech intelligibility, it is desirable to reject the diffuse sound and enhance the direct, anechoic portion of sound. One approach is to apply a varying gain in the time-frequency domain that will successively boost microphone signal when direct sounds dominates and reduces it otherwise, leading to isolation of direct signal. Indeed, even in continuous speech, the direct signal has its energy spread across very few bins, as illustrated in Fig. 9. The speech is said to be sparsely distributed in the time-frequency domain [8].

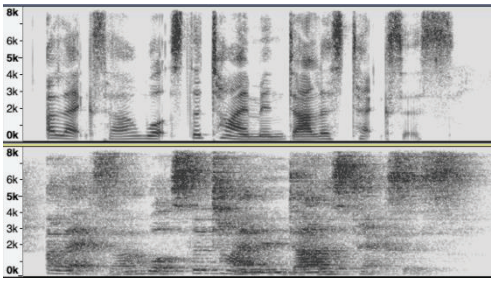


Figure 9. Comparison of spectrograms of anechoic (top) and reverberant speech (bottom). The anechoic speech has its energy spread across less bins than the reverberant ones.

This means that whatever signal is recorded in between the few bins of dominant anechoic speech, it is undesirable and can be theoretically attenuated without distorting speech. It is therefore useful to be able to detect which bins are direct or diffuse dominants to come up with a de-reverberation gain strategy.

A good estimator for this use is the coherence as defined by Thiergart [9]. For a pair of omnis signals $MIC_1(k)$ and $MIC_2(k)$ where k is the wavenumber, it is defined as:

$$\gamma_{12}(kd) = \frac{\Phi_{12}(kd)}{\sqrt{\Phi_{11}(kd)}\sqrt{\Phi_{22}(kd)}}, \qquad (17)$$

where $\Phi_{12}(kd)$ is the power spectral density (PSD) between the two microphones, $\Phi_{11}(kd)$ and $\Phi_{22}(kd)$ are the corresponding auto PSDs, and d is the spacing between microphones.

Using two spaced omnis, the coherence for diffuse sound is frequency dependent: it hovers around 0 at higher frequencies and tends to 1 in lower frequencies. The cut-off frequency becomes lower as the spacing between the microphones is increased. On the contrary, the coherence for direct sounds is 1 at all frequencies as illustrated in Fig. 10.

This estimator is therefore a good discriminator between direct and diffuse sounds at higher frequencies as the coherence values are drastically different for the two types of sounds. This estimator has been used to build de-reverberation algorithms such as the one proposed by Yousefian [10].

As an illustration of the separation capabilities, the histogram of coherence values for purely direct and purely diffuse sound is displayed in Fig. 11.

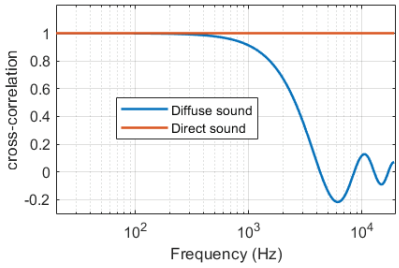


Figure 10. Comparison of coherence for diffuse and direct sounds, using a pair of microphones with 4 cm spacing.

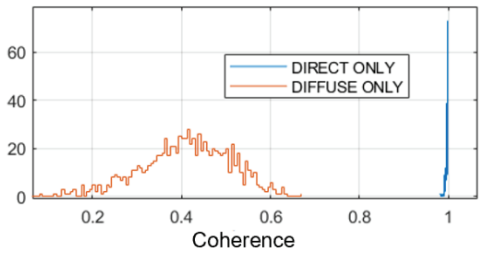


Figure 11. Histogram of measured coherence values between two omni microphones of spacing 4 cm, for direct then diffuse white noise sound (bins computed for 10 ms window, bandpass between 2 kHz to 4 kHz, sampling rate 16 kHz). The distributions are well separated, leading to diffuse-direct classification with high confidence.

However, at lower frequencies the coherence is 1 for both direct and diffuse sounds, leading to difficulties of separating the two.

3. ADVANTAGES OF CO-LOCATED DIRECTIONAL SENSORS SOLUTION

As opposed to omni arrays, an overview of solutions using co-located directional sensors and some of their advantages is presented. In these implementations, the sensors are close together (less than 1 cm of spacing). Therefore, there is not enough time difference between sensors to leverage to figure out directionality with good robustness. On the other hand, the relative amplitude between sensors is the main cue. This leads to major

differences in localization and beamforming methods and performance.

Different hardware implementations have been discussed along the co-located concept: stereo microphones for robot audition [11], ambisonics or sound-field microphone and acoustic vector sensors (AVS) [12]. A few systems are shown in Fig. 12.

Figure 12. Illustration of a couple directional sensors systems: an ambisonics microphone from Sennheiser Ambeo (on the left) and a 3D sound intensity probe from Microflown (on the right).

While omnis now have very flat frequency responses and low noise [13], typical directional microphones have a decreasing sensitivity at lower frequencies as they sense the pressure gradient, making it impractical for the capture of quiet distant sounds with high SNR. Velocity sensors using heated wires also suffers from bandwidth limitations because of diffusion effects [14]. However, recent developments of velocity sensors inspired by insects show promise of high, flat sensitivity dipoles across all audio frequencies [2], displayed in Fig. 13. These acoustic flow sensors would be microphones of choice to pair with an omni to build an AVS and address far field audio use case. The authors presented early results on such applications in [15].

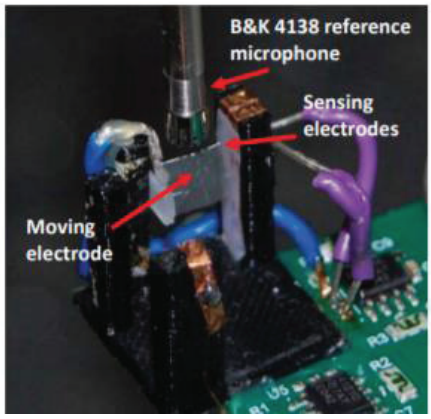


Figure 13. Picture of a flow-sensing prototype from [2].

To illustrate such capabilities, this section is presented using, as an example, an ideal system having a monopole (named W) and two orthogonal velocity sensors (X and Y) with similar frequency responses and sensor noise levels. The system is illustrated in Fig. 14 and the relationship between the microphone signals and the plane waves is expressed in Equation (18):

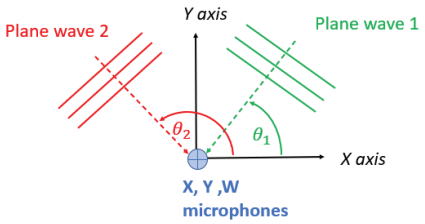


Figure 14. Polar patterns and associated coefficients using a WXY sensor with two incoming plane waves.

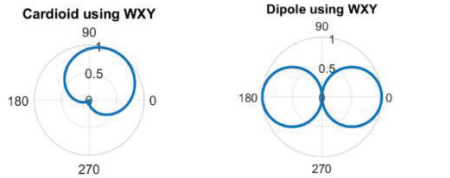
$$\begin{cases} X(t) = \cos(\theta_1).s_1(t) + \cos(\theta_2).s_2(t) + n_X(t), \\ Y(t) = \sin(\theta_1).s_1(t) + \sin(\theta_2).s_2(t) + n_Y(t), \\ W(t) = s_1(t) + s_2(t) + n_W(t). \end{cases} \quad (18)$$

3.1 Beamforming

A system that outputs WXY is convenient to use for beamforming, as any first-order beam (dipoles, variety of cardioids) in the XY plane can be created simply by computing a linear combination of the sensor outputs [16]:

$$b(t) = \alpha_1.W(t) + \alpha_2.X(t) + \alpha_3.Y(t). \quad (19)$$

In Fig. 15, it is illustrated the polar patterns with difference choices of α_i.



Coefficients for cardioid	
α_1	0.5
α_2	0.25
α_3	0.433

Coefficients for sub-cardioid	
α_1	0
α_2	1
α_3	0

Figure 15. Polar patterns and associated coefficients using an ideal WXY sensor.

As the sensor's directivities are constant across all audio frequencies and the coefficients α_i are frequency-independent, so is the output beam, as opposed to omni array. This is illustrated in Fig. 16.

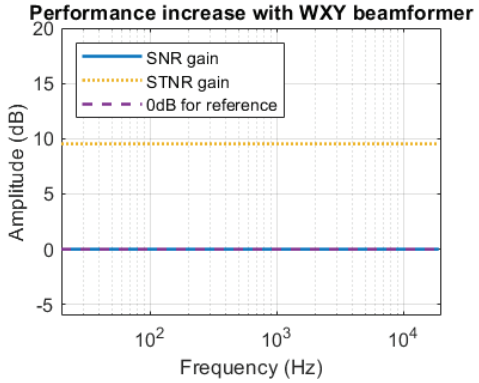


Figure 16. Performance improvement for a dipole using WXY. The SIR gain is not displayed as it is infinite. The STNR is constant at all frequencies.

Additionally, the beamformers using WXY are more robust to microphone mismatch than DMA beamformers, as illustrated in Fig. 17. Indeed, DMA are based on subtraction of pairs, thus perfect cancellation requires the exact same amplitude at the two microphones.

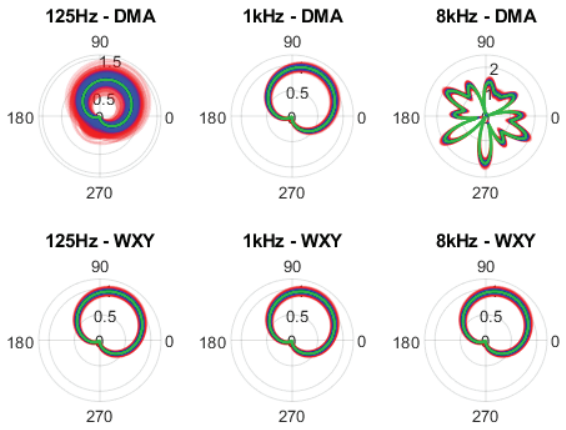


Figure 17. Comparison of polar patterns between DMA and WXY for different frequencies. In green, the curve without mic mismatch. In blue is a superposition of 4000 polar patterns with random mismatch of +/-0.5 dB and +/-2 deg. In red, random polar patterns with +/-1 dB and +/-5 deg mismatch, representative of typical MEMS microphones [12].

3.2 Localization

Similarly, the amplitude difference between microphones can be used to localize sound using WXY. While a few algorithms have been proposed [17], a simple one is the energy-based formula of [18]. We illustrate this for the pair X and Y:

$$AoA_{XY} = \arctan\left(\sqrt{\frac{|Y|^2}{|X|^2}}\right). \qquad (20)$$

In the following, the performance of the AoA estimation using Equation (20) for XY and the GCC method (as in sub-section 2,2) for a pair of omnis in the presence of sensor amplitude and phase mismatch, respectively, is studied. It is assumed that the phase and gain mismatch are taken from a gaussian distribution with a given standard deviation, and look at the standard deviation of the estimated AoA. Results are displayed in Fig. 18. Note that the amplitude mismatch has minor impact on the AoA using omnis, as the shape of the cross-correlation is not influenced by the difference of levels. Reciprocally, the phase mismatch has no effect on the XY AoA as a slight delay between signals do not influence significantly the level difference.

It can be seen that the AoA performance using XY is frequency-independent and signal-independent. On the contrary, the best performance for the omnis is achieved when the peak of the cross-correlation is the sharpest, which occurs for white noise signals and sine waves with higher frequencies. The localization using XY can be also performed at very high frequencies, where the GCC would fail because of spatial aliasing as explained before.

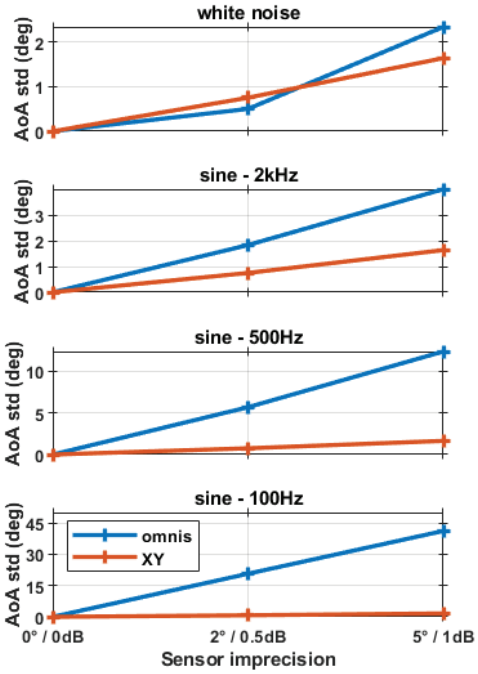


Figure 18. The Localization performance decreases when sensors have phase and gain mismatch (respectively in degrees and dB), in the case of four signals with different frequency content. Phase and gain mismatch are evaluated for both systems.

3.3 Direct-diffuse sound separation

As opposed to omni arrays, the coherence between two orthogonal dipoles tends to 0 at all frequencies in case of diffuse sounds[2], and 1 at all frequencies for direct sounds [8]. Therefore, the estimated coherence helps to separate direct and diffuse sounds with higher confidence, especially at the lower frequencies.

In the graphs of Fig. 19, it can be seen that the coherence of direct sounds in the presence of sensor noise drops below its noiseless value of 1. Direct-diffuse separation can only be achieved if the estimated coherence for direct sounds with sensor noise remain above the coherence value of diffuse noise.

In the bottom graph (mid frequencies), in the case of the XY, the distributions are well separated, while it overlaps for the two omnis. In the top graph (low frequencies), the separation still exists for the XY, but in the case of the omnis the correlation of diffuse sound is actually higher than the measured correlation with direct sound and sensor noise, making the identification between diffuse and direct sound impossible.

[2] The actual measured values are spread around this value because of the finite duration of analysis window.

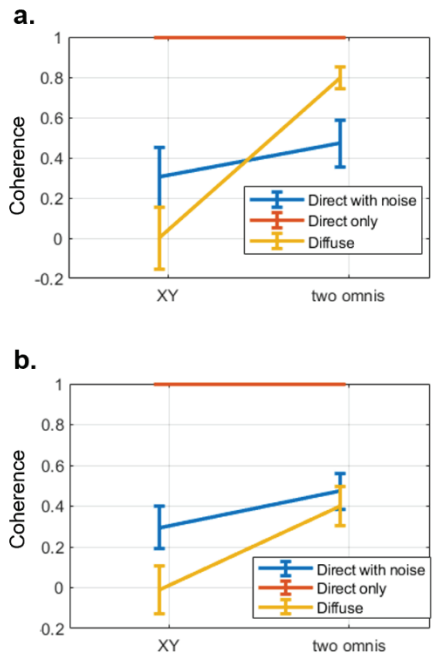


Figure 19. Comparison of the spread of histogram values of the coherence between XY and two omnis in the case of purely diffuse, purely direct, and direct with sensor noise[3]. (a.) White noise sound source band-passed between 125 to 500 Hz. (b.) White noise band-passed between 500 Hz to 2 kHz.

4. CONCLUSIONS

An overview of omnidirectional microphone arrays capabilities was presented then compared to co-located directional microphone arrays. It was shown that directional microphones arrays do not suffer from sensor noise amplification and spatial aliasing, leading to measurable gain in performance in terms of sound localization, beamforming and direct-diffuse sound separation. These promises are driving the design and fabrication of an AVS array using ta new kind of velocity sensors to address far-field audio application.

5. REFERENCES

[1] H.-E. de Bree: "An overview of Microflown Technologies," *Acta Acustica united with Acustica*, pp. 163-172, 2003.

[2] R. N. Miles, M. Farahikia, S. Leahy: "A flow-sensing velocity microphone," *SENSOR 2019 IEEE,* pp. 1-4, 2019.

[3] J. Benesty, J. Chen: *Study and Design of Differential Microphone Arrays,* Springer Topics in Signal Processing, Vol. 6, Springer-Verlag, Berlin, 2013.

[4] I. J. Tashev, *Sound Capture and Processing – Practical Approaches,* John Wiley & Sons, 2009.

[5] M. Brandstein, D. Wards: *Microphone Arrays – Signal Processing Techniques and Applications*, Springer-Verlag, Berlin, 2001.

[6] C. Knapp, G. Carter: "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 24 No. 4,* pp. 320-327, 1976

[7] C. Zhang, D. Florencio, Z. Zhang: "Why does PHAT work well in low noise, reverberative environments?" *ICASSP 2008,* pp. 2565-2568, 2008.

[8] D. Wang: "Time-Frequency Masking for Speech Separation and Its Potential for Hearing Aid Design," *Trends in Amplification,* Vol. 12 No. 4, pp. 332-353, 2008.

[9] O. Thiergart, G. Del Galdo, E. Habets: "On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation", *The Journal of the Acoustical Society of America 132,* pp.2337-2346, 2012.

[10] N. Yousefian, P. C. Loizou: "A Dual-Microphone Speech Enhancement Algorithm Based on the Coherence Function," *IEEE Trans Audio Speech Lang Processing,* Vol. 20 No. 2, p.599-609, 2012.

[11] F. Asanp, M. Morisawa, K. Kaneko, K. Yokoi: "Sound Source Localization using a single-point stereo microphone for robots," *025 APSIPA,* pp.76-85, 2015.

[12] H.-E. de Bree, J. W. Wind: "The acoustic vector sensor: a versatile battlefield acoustics sensor," *Proc. of SPIE,* Vol. 8046, 2011.

[13] Infineon: "IM69D130 – High performance digital XENSIV™ MEMS microphone", 2017

[14] H.-E. de Bree: *The Microflown E-Book,* online, 2009.

[15] L. Carneiro, F. Lepoutre, M. Wang: "Speech localization, classification and separation using a particle velocity sensor", *CIRMMT research workshop on digital signal processing,* 2019.

[16] B. Rafaely: *Fundamentals of Spherical Array Processing,* Springer Topics in Signal Processing Vol. 8, Springer-Verlag, Berlin, 2015.

[17] J. T. Fricke, H.-E. de Bree, A. Siegel, H.-P. Schade: "Source Localization with Acoustic Vector Sensors," *Proc. of the Acoustics High Tatras 2009 – 24th International Acoustic Conference – EAA Symposium,* 2009.

[18] C. A. Dimoulas, K. A. Avdelidis, G. M. Kalliris, G. V. Papanikolaou: "Sound source localization and B-format enhancement using soundfield microphone sets", *AES convention 122,* 2007.

[3] Sensor noise has been added to obtain a SNR of 0 dB.