SOURCE LOCALIZATION USING A SPHERICAL MICROPHONE ARRAY

Madhu Ashok
University of Rochester
mashok@ur.rochester.edu

Erik Nunez
enunez@ur.rochester.edu

Kyle Ohlschlager
kohlschl@u.rochester.edu

ABSTRACT

An algorithm has been devised, and subsequently proposed to process an Eigenmike 32-channel *.wav* file in order to localize *N* sources. Using the angular data of the 32 capsules, sound source locations can be estimated within the acoustic field, with respect to their energy and phase spectrum. We supervise the general energy direction of arrival to choose the set of microphones for triangulation in the forward direction. Given the phase difference between two microphones on the horizontal axis, the angle of arrival – θ can be estimated. Similarly, the vertical angle of arrival - φ can be estimated from the phase difference between two azimuthal microphones. Using this information, a Thetagram and Phigram can be generated, which illustrate angle of arrival with respect to frequency and time. Non-negative matrix factorization was used to preliminarily separate sources in the short-time Fourier transform. We subsequently processed the difference between channels for the angle of arrival for individual dictionaries.

Index Terms - spatial audio, source localization, microphone array processing, NMF

1. INTRODUCTION

Let us consider the simple case of localizing a source between two microphones. In a paper by Chanwoo Kim et. al. [1], source separation was performed on two microphone signals by first finding the relative phase difference between the two channels A and B .

$$\Delta\varphi[fr, \omega_k] = Arg(X_B[fr, \omega_k]) - Arg(X_A[fr, \omega_k]) \quad (1)$$

Where $X_A[fr, \omega_k]$ represents the short-time Fourier transform (STFT) of the signal $x_A[n]$ in channel A, frame – fr, and frequency bin - $\omega_k = \frac{2\pi k}{N}$ (N- FFT length). The phase, $\Delta\varphi[fr, \omega_k]$, was used to calculate the estimated angle of arrival with respect to the central axis between each microphone.

.

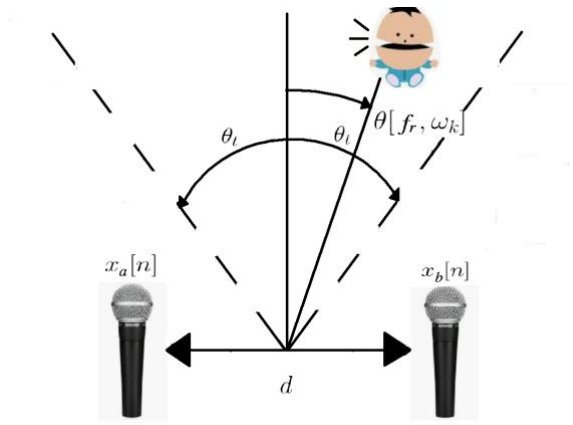


Figure 1. Angle of arrival, $\theta[fr, \omega_k]$, between two microphones A and B separated by a distance - d. The half angle - θ_t is defined from the central axis.

$$\theta[fr, \omega_k] = \sin^{-1}\left(\frac{c\,\Delta\varphi[fr, \omega_k]}{f_s\,\omega_k\,d}\right) \quad (2)$$

Using the speed of sound in air - c, the distance between the microphones - d, the sampling frequency - f_s, phase difference - $\Delta\varphi[fr, \omega_k]$, and the angular frequency - ω_k, the angle of arrival can be estimated.

2. BACKGROUND

The authors propose extrapolating this angle-of-arrival-estimate to the 32 unique angles of capsules on an Eigenmike. [2], The Eigenmike is an array of microphones, or capsules, uniformly distributed about an 8.4cm diameter spherical baffle, capable of effectively receiving audio information within a three dimensional space. Given the angular coordinates of the 32 channels of the Eigenmike [2], we can estimate angle of arrival between any combination of the 32 channels. The next section summarizes our algorithm for a Thetagram, which the authors define to be a visual representation of the angle of arrival with respect to the angular frequency and time. A Thetagram will be the same dimensions as the short-time Fourier transform of the audio signal, but contain angular information of onsets that exceed a threshold in log spectra.

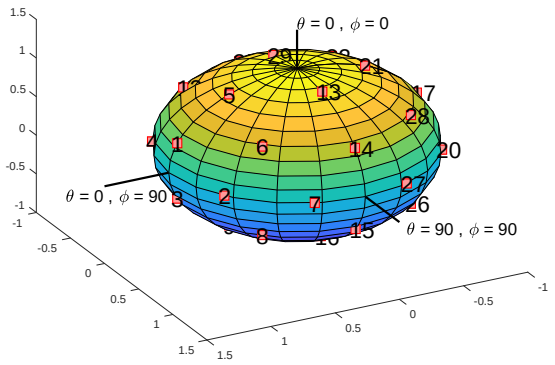


Figure 2. Positions of Eigenmike capsules, where $[\theta = 0°,\ \varphi = 90°]$ is the location of the logo facing the performers.

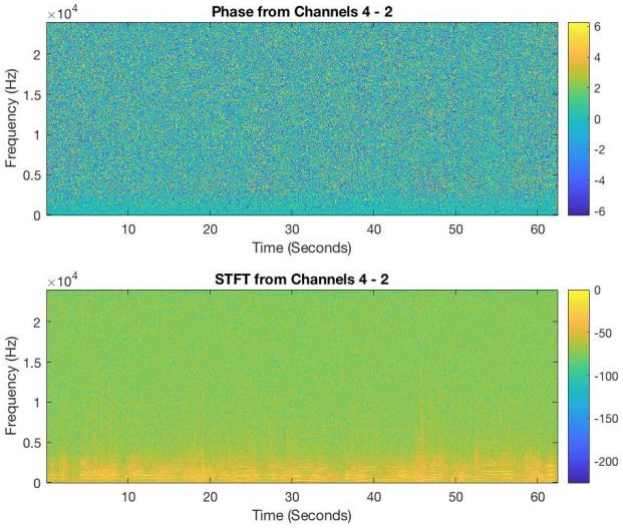


Figure 3. Phase and Spectrogram Differences between Channels 4 & 2 for a recording at Eastman Kodak. There are multiple sources, with the predominant energy coming from a clarinet player towards channel 2 of the Eigenmike.

3. THETAGRAM

Eigenmike *.wav* files comprise of 32 channels of audio at a sampling rate of - f_s and can be manipulated in MATLAB like stereo *.wav* files. An hour of recording on an Eigenmike can correspond to >10 Gb of data, so the audio samples are segmented into frames. We perform the short-time Fourier transform on each frame and channel - i of audio, recording the phase - $Arg(X_i[fr, \omega_k])$ and STFT - $X_i[fr, \omega_k]$. Strategically choosing channels A and B on the Eigenmike can lead to localization along the horizontal axis, vertical axis, or any other triangulation between channels. Figure 3 illustrates localization along the horizontal axis $\pm\theta_t$ between channels 2 and 4, which are located forward-left and forward-right respectively. We define the half angle between capsules to be:

$$\theta_t = \frac{90}{\pi} \tan^{-1}\left(\frac{|\overrightarrow{m_A} \times \overrightarrow{m_B}|}{\overrightarrow{m_A} \cdot \overrightarrow{m_B}}\right) \qquad (3)$$

Where $\overrightarrow{m_A}$ is the vector position of channel A calculated from the capsule orientations. The half angle defines the bounds of angle-of-arrival: $-\theta_t < \theta[fr, \omega_k] < \theta_t$.

Onsets can be seen in the STFT (Figure 3) corresponding to a clarinet player that is located in the back-left of Kodak Hall (Figure 4). The onset strength is determined by the following expression:

$$\zeta_{A,B} = \log(1 + \gamma|X_B[fr, \omega_k] - X_A[fr, \omega_k]|) \qquad (4)$$

Here γ is the compression factor, which can be optimized for detecting onsets in the spectral differences between channels A and B. By assigning a minimum threshold for onsets, $\zeta_{th}\ (dB)$, we can filter the angle of arrival matrix - $\theta[fr, \omega_k]$ - to only display onsets greater than the specified threshold:

$$\Theta_{A,B}[fr, \omega_k] = \{\theta[fr, \omega_k],\ for\ \zeta_{A,B}[fr, \omega_k] > \zeta_{th}\} \qquad (5)$$

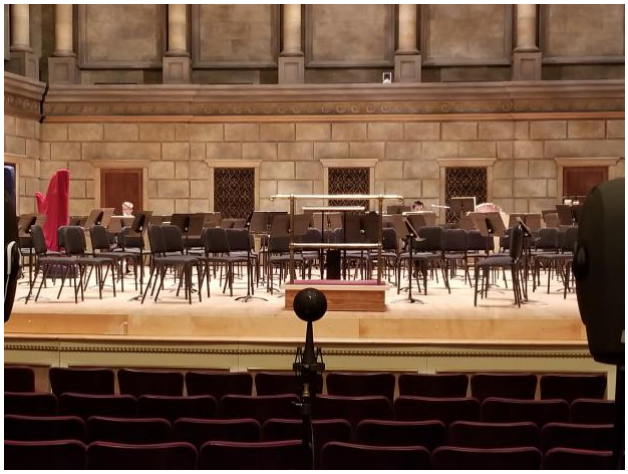


Figure 4. Eigenmike location in Eastman Kodak Hall. There is a clarinet player on the left (near Channel 2) along with noise from other sources in the hall.

When the threshold is set to -55 dB, we obtain the following Thetagram, $\Theta_{A,B}[fr, \omega_k]$, between channels A and B:

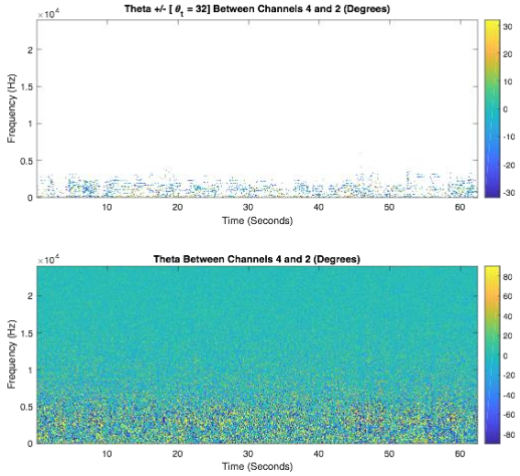


Figure 5. Thetagram (top) generated from the log-spectra difference between channels 4 and 2 (Figure 3) and the angle-of-arrival - $\theta_{2,4}[fr, \omega_k]$ (bottom).

4. TRIANGULATION

In most recording environments the sources are located in a known hemisphere of 4π steradians. The Eigenmike logo is located between channels 1, 2, 3, and 4 at $[\theta = 0°,\ \varphi = 90°]$, which is generally directed towards sources or the center of performers. Professor Ming-Lun Lee organizes spatial audio recordings in Eastman Kodak Hall with the Eigenmike located equidistant from the left and right walls ($\theta = 0°$) and elevated towards the center of the stage, ideally $\varphi = 90°$. Therefore, we can triangulate the sources on stage through channels 5, 7, 9, 11:

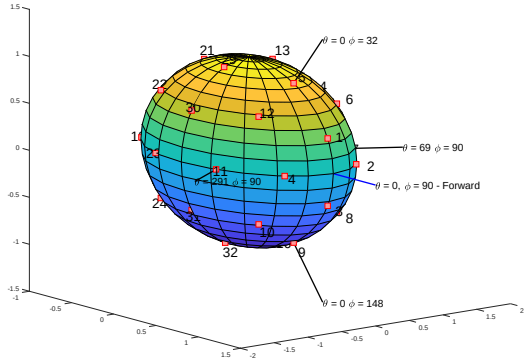


Figure 6. Angle of arrival between channels 7 ($\theta = 69°$) and 11 ($\theta = 291°$) will determine horizontal arrival $0° < \theta < 69°\ and\ 291° < \theta < 360°$. Angle of arrival between channels 5 ($\varphi = 32°$) and 9 ($\varphi = 148°$) will determine vertical arrival $32° < \varphi < 148°$.

Energy direction of arrival - $R_n(\theta, \varphi, \rho)$ can be simplified to six bases: forward, backward, left, right, up, and down. By summing over the unit direction of each microphone capsule - $\overrightarrow{m_i}$ multiplied with the pressure - $x_i[n]$ measured at each microphone capsule - i:

$$R_n(\theta, \varphi, \rho) = \sum_{i=1}^{M} x_i[n] * \overrightarrow{m_i} \tag{6}$$

The energy direction of arrival can be estimated to a general quadrant of 4π steradians by taking the RMS over $R_n(\theta)$ and $R_n(\varphi)$:

$$R_n(\theta) = \tan^{-1} \frac{R_n \sin\varphi \sin\theta}{R_n \sin\varphi \cos\theta} \tag{7}$$

$$R_n(\varphi) = \cos^{-1} \frac{R_n \cos\varphi}{R_n} \tag{8}$$

The clarinet recording in Kodak Hall output $R_{RMS}(\theta) = 2.19°$ and $R_{RMS}(\varphi) = 95.80°$, located between channels 5, 7, 9, and 11 in the forward direction towards the stage. These triangulation channels were chosen strategically in order to determine the horizontal and vertical angles of arrival with respect to the stage, while reducing the effects of diffraction off the sphere. Channels on opposite sides of the Eigenmike will experience the largest effects of diffraction, as the microphone is a solid sphere smaller than the wavelengths of sound.

Given the four channel coordinates for triangulation in the forward direction, a Thetagram and Phigram (Angle of arrival estimate for φ) can be generated for the clarinet recording at Eastman. Channels 7 $[\theta = 69°,\ \varphi = 90°]$ and 11 $[\theta = 291°,\ \varphi = 90°]$ will estimate the angle of arrival for θ, while channels 5 $[\theta = 0°,\ \varphi = 32°]$ and 9 $[\theta = 0°,\ \varphi = 148°]$ will estimate the angle of arrival for φ.

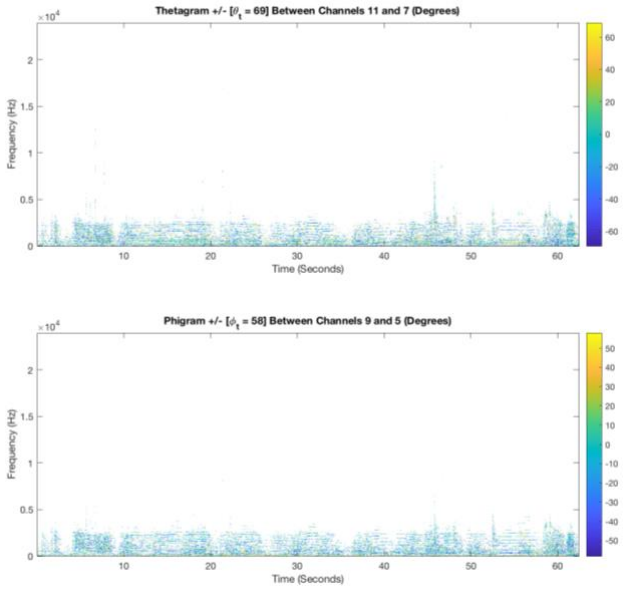
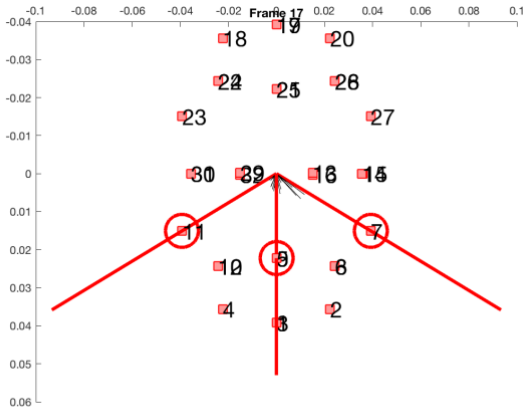


Figure 7. Thetagram for channels 7 and 11 (top) indicating the angle $0° < \theta < 69° \; and \; 291° < \theta < 360°$ of arrival. Phigram for channels 5 and 9 (bottom) indicating the angle $32° < \varphi < 148°$ of arrival.

These figures indicate complexity in the Thetagram and Phigram due to multiple sources and onsets. When the data is simplified to a single frame (Figure 8), various distributions can be seen in horizontal and vertical angles of arrival. The authors postulate that the peak log-spectra differences (greatest magnitude for black and blue lines) indicate the direction of arrival of a source. In the top figure, two clear distributions can be observed indicating two possible sources in the frame.



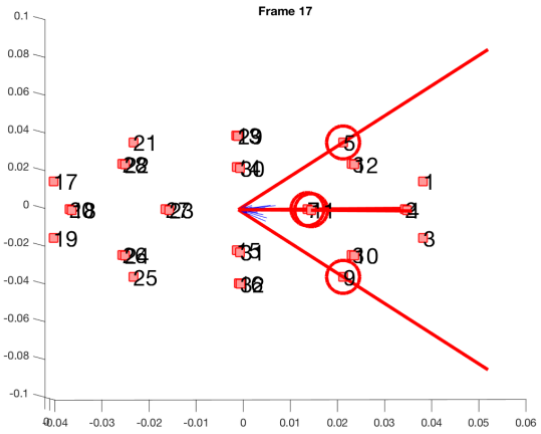


Figure 8. Triangulation of arrival for a single frame (17), with black lines indicating the horizontal angle of arrival – θ (top) and blue lines for vertical angle of arrival - φ (bottom). The magnitudes of the black and blue vectors are the log spectra difference from the channel pairs – $\zeta_{7,11}[17, \omega_k]$ $(black)$ and $\zeta_{5,9}[17, \omega_k]$ $(blue)$.

5. NMF FOR SOURCE IDENTIFICATION

The Thetagram and Phigram show fundamental frequencies of the clarinet with various harmonic content arriving at angles $\theta \; and \; \varphi$. Naively choosing a range of angles of arrival and filtering the Thetagram/Phigram will cut out harmonics arriving in directions that are not from the source. The harmonic content of a note on the clarinet may arrive from multiple reflections in the room, thus various angles of arrival. To combat this issue, the authors propose using Non-negative matrix factorization to separate the STFT into dictionaries of notes. Non-negative matrix factorization has been used to separate sources by generating a dictionary - $W(:, r)$ for r columns (features) multiplied with activations - $H(r, :)$ from the corresponding dictionary [6].

$$H_{l+1} = H_l \odot \left(\frac{W_l^{\mathsf{T}} V}{W_l^{\mathsf{T}} W_l H_l} \right) \tag{9}$$

$$W_{l+1} = W_l \odot \left(\frac{V H_{l+1}^{\mathsf{T}}}{W_l H_{l+1} H_{l+1}^{\mathsf{T}}} \right) \tag{10}$$

Given an input vector of the STFT difference between channels $A \; and \; B$:

$$V = abs(X_B[fr, \omega_k] - X_A[fr, \omega_k]) \tag{11}$$

Dictionaries (Equation 9) and activations (Equation 10) iteratively update $(l \rightarrow l + 1)$ vectors until their estimated output vector: $V_{est} = W_{l+1} H_{l+1}$ converges to V. We can measure the KL divergence between $V \; and \; V_{est}$ through the Frobenius norm of the following expression [7].

$$KL_{l+1} = \left\| \left(V \odot \log\left(\frac{V}{V_{est}} \right) \right) - V + V_{est} \right\| \tag{12}$$

The divergence is calculated for each iteration until $V_{est} \approx V$. After 50 iterations and $r = 20$ features, the following dictionaries and activations were created for the clarinet audio file:

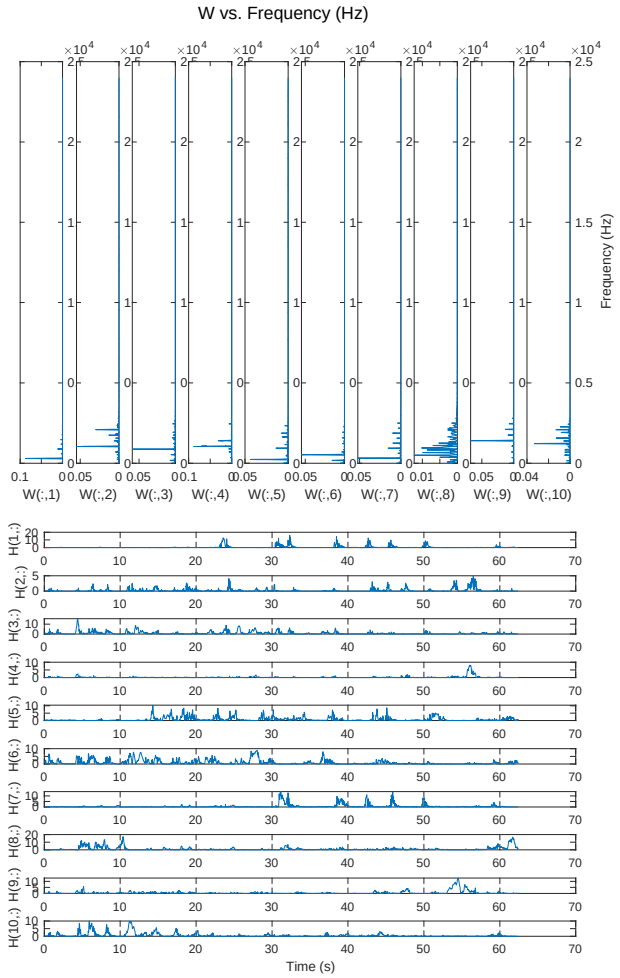


Figure 9. The first 10 dictionaries - $W(:, 1-10)$ shown on the top figure, and the first 10 activations - $H(1-10,:)$ on the bottom figure.

By multiplying the tenth dictionary $W(:,10)$ and activation $H(10,:)$, an estimated spectrogram $V_{10} = W(:,10)H(10,:)$ can be used as an input to the Thetagram/Phigram:

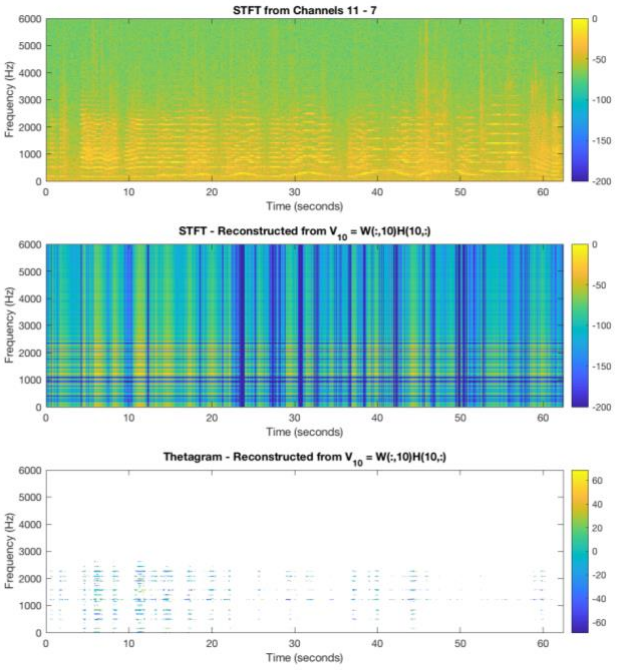


Figure 10. The short-time-Fourier-transform of the tenth dictionary reconstructed - $V_{10} = W(:,10)H(10,:)$ (middle graph). The reconstructed STFT is fed as an input to the Thetagram (bottom graph).

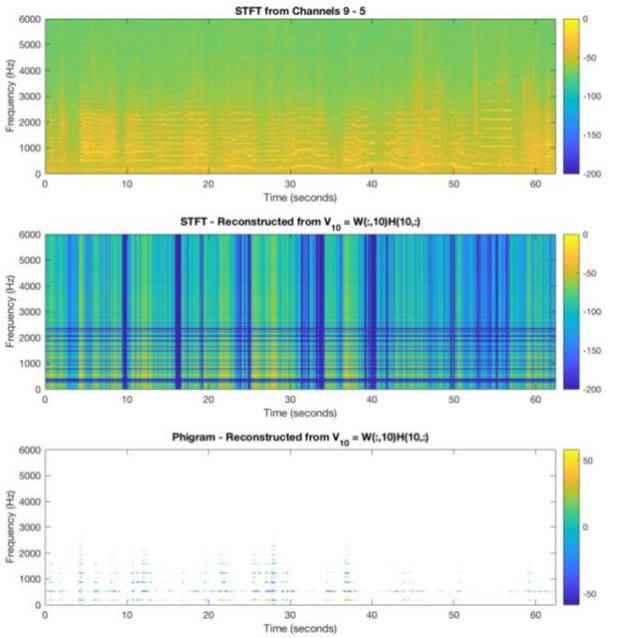


Figure 11. The short-time-Fourier-transform of the tenth dictionary reconstructed - $V_{10} = W(:,10)H(10,:)$ (middle graph). The reconstructed STFT is fed as an input to the Phigram (bottom).

6. FUTURE WORK

The authors' research and implementations can be applied to a myriad of applications, including source separation for speech diarization, automatic audio transcription, acoustic monitoring in various environments, and adaptive array processing using minimum variance in channels. The benefit of using the

multichannel Eigenmike with the authors' proposed algorithms is it lends to greater versatility and robustness in data collection, which theoretically increases with each channel used for audio data retrieval and processing.

The current direction that we would like to move forward within algorithm development is classifying microphones that are closest to each source with an adaptive minimum variance method. Once the sources are separated, we can create spectral weights for each source and determine the number of sources through what is left after each successive classification and filtering of a source.

Advances in bioacoustic technology have allowed for the development of wireless recording arrays that permit ambient recordings at multiple locations over long periods of time. With these sort of systems now in development, we can choose the direction of research to follow in the classification of what is considered noise or for example, population analysis of bird species that are on the brink of extinction. With the rates of how some species are going extinct, it's is imperative to have accurate population numbers which can tedious to send out teams of researchers to take data samples on a regular consistent basis. Recently technology has been benefiting from the rise of cheaper data storage and increase in processing power available to the average consumer, The same bird environment datasets could be used to develop better sound source cancellation signal processing methods. Long-term feature vector analysis of environments is a relatively new field that could combine bioacoustic technology and machine learning.

7. CONCLUSION

The devised algorithm in conjunction with the 32-channel Eigenmike proved effective in sound source localization, provided even an acoustically complex environment such as the Eastman Kodak Hall, where reverberations and reflections are prevalent, and the source distances from the Eigenmike, or one sources position relative to another source, may be very near, or quite far, respectively, lending to a complicated setting for audio signal analyzation and processing. While the Eigenmike we used has been designed with a considerably small 8.4cm diameter spherical baffle, a concern was that corresponding capsules were distributed too near to one another; close enough to cause difficulty in effectively detecting phase differences between incoming audio signals. As observed, however, analyzation of the energy spectra and phase difference between any pair of the thirty-two microphones yielded distinguishable results, such that the formerly mentioned concern was of no apparent issue. Using the authors' algorithm, in conjunction with the 32-channel Eigenmike, sound source localization was successful, and paves a concrete foundation for future implementation in sound source separation, even in acoustically diverse, and complex environments.

8. REFERENCES

[1] C. Kim and K. K. Chin, "Sound source separation algorithm using phase difference and angle distribution modeling near the target," in INTERSPEECH-2015, Sept. 2015, pp. 751–755.

[2] mh acoustics staff, em32 Eigenmike microphone array release notes (v17.0), mh acoustics, 2013.

[3] H. Sun, H. Teutsch, E. Mabande and W. Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPRIT with spherical microphone arrays," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, 2011, pp. 117-120. doi: 10.1109/ICASSP.2011.5946342

[4] D. Malioutov, M. Cetin and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," in *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010-3022, Aug. 2005. doi: 10.1109/TSP.2005.850882

[5] Park, Munhum & Rafaely, Boaz. (2005). Sound-field analysis by plane-wave decomposition using spherical microphone array. The Journal of the Acoustical Society of America. 118. 3094-3103. 10.1121/1.2063108.

[6] Müller, M. (2015). *Fundamentals of music processing: Audio, analysis, algorithms, applications*.

[7] Smaragdis, P., & Brown, J. C. (2003, October). Non-negative matrix factorization for polyphonic music transcription. In *IEEE workshop on applications of signal processing to audio and acoustics* (Vol. 3, No. 3, pp. 177-180).