A NOVEL TIME-DELAY-OF-ARRIVAL ESTIMATION TECHNIQUE FOR MULTI-MICROPHONE AUDIO PROCESSING

Jes Thyssen, Ashutosh Pandey, and Bengt Jonas Borgström

Broadcom Corporation
Irvine, California, USA

ABSTRACT

Multi-microphone speech enhancement requires knowledge of relative Time Delay of Arrival (TDOA) of the desired acoustic source at microphones. This paper presents a novel TDOA estimation method, Steered Null Error PHAse Transform (SNE-PHAT), which exploits null-steering to improve estimation robustness. The method is formulated to be computationally efficient. A generalization to provide frequency-dependent TDOA estimates is proposed. Experimental results demonstrate that SNE-PHAT outperforms the Generalized Cross Correlation PHAse Transform (GCC-PHAT) method, particularly in the presence of background noise. Additionally, experiments illustrate the benefits of using frequency-dependent TDOA estimation.

Index Terms— Sound source tracking, localization, null-steering, beamforming, microphone array, GCC-PHAT, TDOA.

1. INTRODUCTION

The use of mobile handsets in hands-free mode requires high quality signals to be transmitted to the far end, even in challenging acoustic conditions. Multi-microphone processing has become popular for noise suppression in mobile phones [1]. Figure 1 shows a simplified block diagram of typical multi-microphone processing for speech enhancement. Microphone signals $x_i(n), i = 1 \dots M$ capture delayed versions of the desired signal $s(n)$, desired signal reflections, noise components, and acoustic echo from the far end. Captured signals are typically transformed into the subband or frequency domain in order to provide spectral resolution during succeeding processing [1]. Beamforming can then be applied to extract an enhanced signal $\hat{s}(n)$. Examples of beamforming methods include the generalized-side-lobe canceller [2] with single-channel post-processing [3]. Note that if the communication system includes playback of a far-end signal, acoustic echo cancellation (AEC) can be applied at various points in the processing chain [4].

Beamforming typically uses relative delay between microphone pairs [5]. A primary microphone is chosen and relative time-delay-of-arrival (TDOA) values between the primary microphone and supporting microphones, τ_{1j} for $j = 2 \dots M$, are estimated. The relative delay values between microphone pairs depend on the angle of incidence of the desired source, θ, with respect to the microphone array, which must be estimated on a frame-by-frame basis. In most systems, the relative delay is estimated as a two-step process. First, raw frame-specific TDOA values are obtained. These estimates are then refined by post-processing using for example, particle-filtering, clustering, mixture-modeling, and Kalman filtering [5, 6, 7, 8].

Accurate estimation of the raw TDOA values is vital for subsequent blocks. Many successful time-delay estimation methods exist,

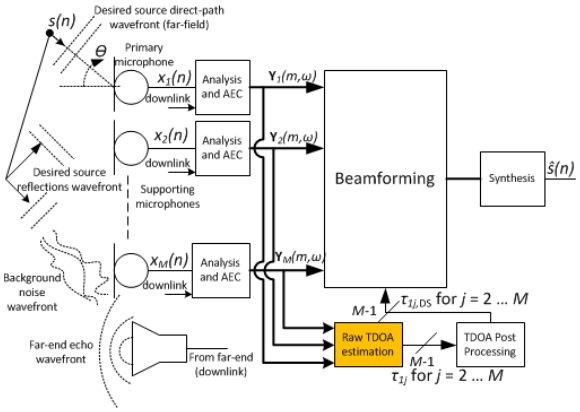

Fig. 1. Block diagram of typical multi-microphone processing for speech communication.

including adaptive eigenvalue decomposition [9], cross-correlation-based methods [10], and maximum-likelihood-based methods [11]. Of these, the Generalized Cross Correlation PHAse Transform (GCC-PHAT) method has proven itself popular due to its computational efficiency and good performance [1, 5, 6]. The GCC-PHAT method estimates TDOAs by finding the angle that maximizes the steered response of a microphone pair over the desired frequency range. The performance of the GCC-PHAT method, however, degrades in the presence of acoustic noise.

This paper presents the Steered Null Error PHAse Transform (SNE-PHAT), a novel method for TDOA estimation that uses null-steering to find the desired time-difference. For the case of two microphone signals, the method determines the delay that maximizes the prediction gain, when predicting one microphone signal from the other microphone signal under optimal (frequency-dependent) gain. Equivalently, SNE-PHAT minimizes the associated prediction error. In addition, the paper provides a mathematically equivalent, yet computationally efficient formulation of SNE-PHAT, that is suitable for low complexity implementation.

A generalization to provide frequency-dependent TDOA estimates, in order to account for variations in TDOA with respect to frequency, is developed. The TDOA for a point source, observed in free far-field conditions, is well-known to be constant for all frequencies. However, real-world acoustic data may exhibit frequency-dependent TDOA due to various acoustic phenomena, and it may prove beneficial for subsequent processing blocks to represent this characteristic. Furthermore, the paper discusses how to address the ambiguity due to spatial aliasing at higher frequencies.

The paper is organized as follows. Section 2 provides a brief summary of the classical GCC-PHAT method and introduces no-

tation. The proposed SNE-PHAT method is presented in detail in Section 3. Experimental results are provided in Section 4, and conclusions are given in Section 5.

2. NOTATIONS AND GCC-PHAT REFERENCE METHOD

The widely-used GCC-PHAT method serves as a reference for the proposed method. Both methods are treated in the frequency-domain, and both are implemented using identical windowing and Fast Fourier Transform (FFT). The time-domain microphone signal $y_i(n)$ for microphone i is represented by $Y_i(m,\omega_k)$, where m denotes the frame number, and $\omega_k = 2\pi k/K$, for $k = 0, 1, \ldots, K/2$, denote discrete frequency points. In the following, the frame index m and subscript k for ω have been omitted to simplify notation, and the microphone signal in the frequency domain is expressed as $Y_i(\omega)$.

Consistent with [10], GCC-PHAT estimates the TDOA τ_{ij} for microphone pair i and j by maximizing the steered normalized cross-correlation function according to

$$\tau_{ij} = \underset{\tau\in\left[-\frac{d_{ij}}{c},\frac{d_{ij}}{c}\right]}{\operatorname{argmax}} \sum_{\omega} \frac{\mathrm{E}\left[Y_i(\omega)Y_j^*(\omega)\right] e^{-j\omega\tau}}{\left|\mathrm{E}\left[Y_i(\omega)Y_j^*(\omega)\right]\right|} \tag{1}$$

where d_{ij} is the distance between microphone i and j, c is the speed of sound in air, $*$ is the conjugate operator and E[.] is the expectation operator. In practice, the expectation operator is implemented with a running mean and serves to reduce the variance of the spectrum estimates.

3. STEERED NULL ERROR PHASE TRANSFORM (SNE-PHAT) METHOD

The GCC-PHAT method described in Section 2 implicitly determines the TDOA by looking in different directions and selecting the direction that maximizes the cross-correlation energy. On the other hand, the method proposed in this paper scans for time delays between a microphone pair i and j and selects the TDOA that minimizes a steered null-error, or as conceptually outlined above, selects the TDOA that maximizes the prediction gain when predicting one microphone signal from the other under optimal (frequency-dependent) gain. For a particular frequency bin, ω, and candidate TDOA, τ, the prediction error is

$$E_{ij}(\omega,\tau) = Y_j(\omega) - G_{ij}(\omega,\tau)e^{j\omega\tau}Y_i(\omega) \tag{2}$$

where $G_{ij}(\omega,\tau)$ is the optimal gain. The corresponding time domain prediction error is denoted $e_{ij}(n,\tau)$. Clearly the optimal gain should not represent any delay (as that is to be captured and represented by the TDOA, τ), and hence, when deriving a solution for the optimal gain, the solution must be constrained to be real and positive. The core cost function for calculating the optimal real gain is formulated as

$$\begin{aligned} C_G(\omega,\tau) =& E\left[E_{ij}(\omega,\tau)E_{ij}^*(\omega,\tau)\right] \\ =& E\left[Y_j(\omega)Y_j^*(\omega)\right] + G_{ij}^2(\omega,\tau)E\left[Y_i(\omega)Y_i^*(\omega)\right] \\ & - 2G_{ij}(\omega,\tau)\mathrm{Re}\left(e^{-j\omega\tau}E\left[Y_j(\omega)Y_i^*(\omega)\right]\right) \\ =& R_{jj}(\omega) + G_{ij}^2(\omega,\tau)R_{ii}(\omega) \\ & - 2G_{ij}(\omega,\tau)\mathrm{Re}\left(e^{-j\omega\tau}R_{ji}(\omega)\right) \end{aligned} \tag{3}$$

where $R_{ji}(\omega) = E\left[Y_j(\omega)Y_i^*(\omega)\right]$, and imposing $G_{ij}(\omega,\tau) = G_{ij}^*(\omega,\tau)$ enforces the gain to be real. The optimal *positive* real gain is found as a constrained optimization according to

$$G_{ij}(\omega,\tau) = \underset{G}{\operatorname{argmin}}\left\{C_G(\omega,\tau)\right\} \text{ subject to } G \geq 0 \tag{4}$$

using the Kuhn-Tucker method [12]:

$$G_{ij}(\omega,\tau) = \begin{cases} 0 & R_{ji}^{\text{dir}}(\omega,\tau) < 0 \\ \frac{R_{ji}^{\text{dir}}(\omega,\tau)}{R_{ii}(\omega)} & R_{ji}^{\text{dir}}(\omega,\tau) \geq 0 \end{cases} \tag{5}$$

where $R_{ji}^{\text{dir}}(\omega,\tau) = \mathrm{Re}\left(e^{-j\omega\tau}R_{ji}(\omega)\right)$.

3.1. Fullband Steered Null Error (SNE)

The fullband prediction gain, $P_{\text{fb}}(\tau)$, of microphone signal j, $y_j(n)$, from microphone signal i, $y_i(n)$, is given by

$$\begin{aligned} P_{\text{fb}}(\tau) =& 10\log_{10}\left(\frac{E\left[y_j^2(n)\right]}{E\left[e_{ij}^2(n)\right]}\right) \\ =& 10\log_{10}\left(\frac{E\left[\sum_\omega Y_j(\omega)Y_j^*(\omega)\right]}{E\left[\sum_\omega E_{ij}(\omega,\tau)E_{ij}^*(\omega,\tau)\right]}\right) \\ =& 10\log_{10}\left(\frac{\sum_\omega E\left[Y_j(\omega)Y_j^*(\omega)\right]}{\sum_\omega C_G(\omega,\tau)}\right), \end{aligned} \tag{6}$$

where $C_G(\omega,\tau)$ is given by Eq. 3 with $G_{ij}(\omega,\tau)$ given by Eq. 5. Technically, the fullband TDOA is found as the delay that maximizes the prediction gain:

$$\tau_{ij}^{\text{fb}} = \underset{\tau\in\left[-\frac{d_{ij}}{c},\frac{d_{ij}}{c}\right]}{\operatorname{argmax}} P_{\text{fb}}(\tau) \tag{7}$$

However, since $\log_{10}(x)$ is a monotonically increasing function, $1/x$ is a monotonically decreasing function, and $E\left[Y_j(\omega)Y_j^*(\omega)\right]$ is independent of τ, this is equivalent to

$$\tau_{ij}^{\text{fb}} = \underset{\tau\in\left[-\frac{d_{ij}}{c},\frac{d_{ij}}{c}\right]}{\operatorname{argmin}} \sum_{\omega} G_{ij}^2(\omega,\tau)R_{ii}(\omega) - 2G_{ij}(\omega,\tau)R_{ji}^{\text{dir}}(\omega,\tau). \tag{8}$$

3.2. Fullband SNE-PHAT

Low-frequency content often dominates speech signals, and at low frequencies (long wavelength) the spatial resolution is poor, resulting in a poorly-defined peak in the cost function expressed by Eq. 8. Hence, it is advantageous to equalize the spectral envelope to some degree in order to provide greater weight to frequencies where the peak of the cost function is more clearly defined. Including such equalization to the SNE results in the SNE-PHAT method. If the microphone spectra are normalized by their magnitude spectrum, then the equalized cross spectra is given by

$$R_{ji}^{\text{eq}}(\omega) = \frac{R_{ji}(\omega)}{\sqrt{R_{jj}(\omega)R_{ii}(\omega)}}, \tag{9}$$

with the power spectra being a special case: $R_{ii}^{\text{eq}}(\omega) = 1$. Basing the calculation of the optimal gain on the normalized cross and power spectra results in the following optimal gain

$$G_{ij}^{\text{eq}}(\omega,\tau) = \begin{cases} 0 & R_{ji}^{\text{dir,eq}}(\omega,\tau) < 0 \\ R_{ji}^{\text{dir,eq}}(\omega,\tau) & R_{ji}^{\text{dir,eq}}(\omega,\tau) \geq 0 \end{cases} \tag{10}$$

where $R_{ji}^{\text{dir,eq}}(\omega,\tau) = \mathrm{Re}\left(e^{-j\omega\tau}R_{ji}^{eq}(\omega)\right)$. The fullband TDOA, corresponding to Eq. 8, becomes

$$\begin{aligned} \tau_{ij}^{\text{fb}} =& \underset{\tau\in\left[-\frac{d_{ij}}{c},\frac{d_{ij}}{c}\right]}{\operatorname{argmin}} \sum_{\omega} G_{ij}^{\text{eq}}(\omega,\tau)\left(G_{ij}^{\text{eq}}(\omega,\tau) - 2R_{ji}^{\text{dir,eq}}(\omega,\tau)\right) \\ =& \underset{\tau\in\left[-\frac{d_{ij}}{c},\frac{d_{ij}}{c}\right]}{\operatorname{argmin}} \sum_{\omega} C_{\text{TDOA}}(\omega,\tau), \end{aligned} \tag{11}$$

where

$$C_{\text{TDOA}}(\omega,\tau) = G_{ij}^{\text{eq}}(\omega,\tau)\left(G_{ij}^{\text{eq}}(\omega,\tau) - 2R_{ji}^{\text{dir,eq}}(\omega,\tau)\right)$$

$$= \begin{cases} 0 & R_{ji}^{\text{dir,eq}}(\omega,\tau) < 0 \\ -\left(R_{ji}^{\text{dir,eq}}(\omega,\tau)\right)^2 & R_{ji}^{\text{dir,eq}}(\omega,\tau) \geq 0 \end{cases} \quad (12)$$

is defined as the frequency-dependent cost function at frequency, ω, and TDOA, τ, and

$$C_{\text{TDOA}}^{\text{fb}}(\tau) = \sum_{\omega} C_{\text{TDOA}}(\omega,\tau) \quad (13)$$

is the fullband cost function.

It should be noted that the normalization of Eq. 9 is identical to that of GCC-PHAT, see Eq. 1, if the expectation operator is omitted. While the expectation is not strictly necessary, it is beneficial as it reduces the variance of the spectrum estimates. It can been implemented as a simple 1^{st}-order running mean with high leakage in order not to compromise tracking of moving sources.

3.3. Frequency-Dependent SNE-PHAT

The unconstrained frequency-dependent TDOA follows directly from Eq. 11 by searching the frequency-dependent cost function according to

$$\tau_{ij}(\omega) = \underset{\tau \in \left[-\frac{d_{ij}}{c}, \frac{d_{ij}}{c}\right]}{\text{argmin}} C_{\text{TDOA}}(\omega,\tau). \quad (14)$$

A better estimate of the true underlying frequency-dependent TDOA can be achieved by constraining the search to a vicinity of the fullband TDOA, for example to a fixed range of $\pm\delta$. Additionally, spatial aliasing results in false peaks (side lobes) in the cost function, Eq. 12, at

$$\tau(\omega) = \tau_{\text{true}} \pm k\frac{2\pi}{\omega}, k = 1, 2, ... \quad (15)$$

and it is advantageous to constrain the search and exclude false peaks from consideration, that is strict the search range to be within a fraction $0 < K < 1$ of the first lobe of spatial aliasing in either direction. Combining the two constraints to the tighter of the two is beneficial because: 1. The spatial aliasing constraint becomes unconstrained at low frequencies, 2. The fixed constraints is not sufficiently constrained at high frequencies to prevent searching the side lobes. The constrained frequency-dependent TDOA is found according to

$$\tau_{ij}(\omega) = \underset{\tau \in \left[\tau_{ij}^{\text{fb}} - \Delta_\tau(\omega), \tau_{ij}^{\text{fb}} + \Delta_\tau(\omega)\right]}{\text{argmin}} C_{\text{TDOA}}(\omega,\tau), \quad (16)$$

where $\Delta_\tau(\omega) = \min\left\{\delta, K\frac{2\pi}{\omega}\right\}$. It should be noted that GCC-PHAT can be extended in a similar manner to provide frequency-dependent TDOA.

Figure 2 provides a qualitative example of the inner workings of fullband and frequency-dependent SNE-PHAT TDOA estimation. The figure provides an illustration of how the constrained frequency-dependent TDOA prevents the selection of a TDOA corresponding to side lobes from spatial aliasing.

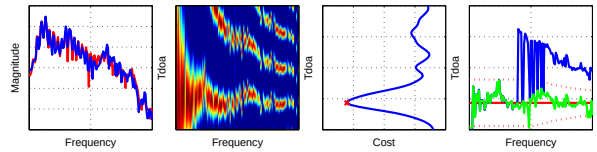

Fig. 2. Example of SNE-PHAT. Left to right: a. Magnitude spectrum of the microphone signals, b. Cost function for frequency-dependent TDOA, $C_{\text{TDOA}}(\omega,\tau)$, c. Cost function for fullband TDOA, $C_{\text{TDOA}}^{\text{fb}}(\tau)$, in blue, fullband TDOA according to Eq. 11 at red cross, d. Fullband TDOA in solid red, unconstrained frequency-dependent TDOA in blue, frequency-dependent TDOA constrained according to Eq. 16 in green (with parameters $\delta = 1.5$ samples, $K = 0.4$), and the TDOA constraints in dotted red.

3.4. Computational Complexity Considerations

Besides the performance, the computational complexity is an important attribute. Expanding the expression for $R_{ji}^{\text{dir,eq}}(\omega,\tau)$, calculating it in pairs of TDOA candidates, $\pm\tau$, leads to

$$R_{ji}^{\text{dir,eq}}(\omega,\pm\tau) = \text{Re}\left(e^{\mp j\omega\tau} R_{ji}^{\text{eq}}(\omega)\right) = \cos(\omega\tau)\frac{\text{Re}\left(R_{ji}(\omega)\right)}{\sqrt{R_{jj}(\omega)R_{ii}(\omega)}} \pm \sin(\omega\tau)\frac{\text{Im}\left(R_{ji}(\omega)\right)}{\sqrt{R_{jj}(\omega)R_{ii}(\omega)}}, \quad (17)$$

where $R_{jj}(\omega)$ and $R_{ii}(\omega)$ are real. Calculating $R_{ji}^{\text{dir,eq}}(\omega,\tau)$ takes:

- 1 multiply, 1 square-root, and 1 division per frequency bin for the inverse numerator (independent of the TDOA candidate),
- 2 multiplies per frequency bin to normalize the real and imaginary parts of $R_{ji}(\omega)$ (independent of the TDOA candidate),
- 2 multiplies and 1 add per frequency bin per *one half* of TDOA candidates, and 1 add per frequency bin per *the other half* of TDOA candidates, to combine the cosine and sine.

From Eq. 12, $C_{\text{TDOA}}(\omega,\tau)$ can be calculated by 1 multiply per frequency bin per TDOA candidate, disregarding the minus by maximizing instead of minimizing. Calculating $C_{\text{TDOA}}^{\text{fb}}(\tau)$ from $C_{\text{TDOA}}(\omega,\tau)$ takes 1 add per frequency bin per TDOA candidate. A similar account for GCC-PHAT of Eq. 1, exploiting similar tricks, leads to the results shown in Table 1. Example parameter settings

Table 1. Complexity of GCC-PHAT and SNE-PHAT

	Div	Sqrt	Mult	Add	WMOPS
Weight	20	20	1	1	
GCC-PHAT	103	103	6592	12669	2.34
SNE-PHAT	103	103	12875	12669	2.97

for Table 1 comprise calculation over 61 candidate TDOAs, using 103 frequency bins (from a 256-point FFT), every 10 millisecond (ms), with complexity weights as indicated in the table to get final WMOPS (Weighted Million Operations Per Second) estimates. Overhead for calculation of spectra, sine, and cosine was omitted. Although SNE-PHAT is approximately 25% more complex than GCC-PHAT, both take up a small portion of what a reasonable budget for a complete multi-microphone algorithm would be.

Since $C_{\text{TDOA}}(\omega,\tau)$ is calculated as part of calculating $C_{\text{TDOA}}^{\text{fb}}(\tau)$, the constrained search of Eq. 16 is the only remaining task to carry out to obtain the frequency-dependent TDOA. Hence, the frequency-dependent TDOA can be obtained at a very small overhead.

4. EXPERIMENTAL RESULTS

In order to demonstrate the proposed methods, experiments were carried out with the TDOA estimation module of a complete cellular speakerphone mode system. Short-time spectral analysis was implemented in the frequency domain with sampling frequency $f_s = 8000$ Hz, analysis window length $L = 160$ samples, window shift $R = 80$ samples, and FFT size $K = 256$ samples. The TDOA estimation used two microphones, $M = 2$, with $d_{12} = 120$ mm spacing. Real-world acoustic data was collected with a 14-microphone mock-up handset shown in Figure 3. The noise robustness of the proposed SNE-PHAT method is first compared to the GCC-PHAT method in terms of fullband TDOA estimation accuracy. Spectrally flat directional desired sources were created at increments of 1^{o} angle-of-incidence. Spherical diffuse noise was generated as described in [13] and was added at different SNRs to the desired source. The source tracking algorithms (SNE-PHAT

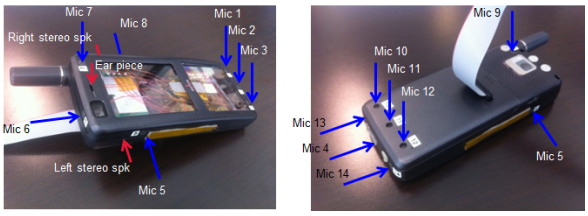

Fig. 3. 14 microphone mock up handset for data collection.

and GCC-PHAT) were used to estimate fullband TDOA values, and estimation error statistics were calculated in the form of mean and variance of absolute error, combined in groups of 30 degrees as shown in Table 2.

Table 2. Average absolute TDOA estimation error

True Angle	SNR (dB)	Absolute error: mean $\pm$ variance	
		GCC-PHAT	SNE-PHAT
[0, 29]	0	5.11 ± 16.15	3.97 ± 8.34
	10	2.45 ± 8.01	1.64 ± 3.21
	20	2.11 ± 6.81	1.57 ± 2.68
[30, 59]	0	1.20 ± 0.85	0.54 ± 0.23
	10	0.44 ± 0.11	0.29 ± 0.07
	20	0.37 ± 0.05	0.28 ± 0.07
[60, 89]	0	0.86 ± 0.45	0.38 ± 0.07
	10	0.33 ± 0.06	0.24 ± 0.02
	20	0.27 ± 0.04	0.23 ± 0.01

As expected, estimation error increases with decreasing SNR and with the angle of incidence ranging from broadside to end-fire. The experimental results demonstrate that SNE-PHAT reduces both the mean and variance of the absolute error, relative to GCC-PHAT, particularly for low SNR conditions. This is advantageous as more accurate raw TDOA estimates benefit subsequent processing in a multi-microphone noise suppression system.

Table 3. BM suppression of DS with frequency-dependent TDOA.

SNR (dB)	Suppression of DS (dB)			
	Diffuse (spherical) noise		Interfering talker	
	Fullband	Freq. dep.	Fullband	Freq. dep.
50	16.4	29.8	14.0	27.3
20	14.3	22.1	13.5	23.4
10	12.8	17.2	11.8	18.0

The performance of frequency-dependent TDOA estimation is demonstrated with dual-microphone signals where a primary talker is the desired source (DS) and spherical diffuse noise or an interfering talker at various levels is the interfering source (IS). The performance was quantified by measuring the suppression of the DS by a blocking matrix (BM) [2]. The BM was adaptive with the phase being determined from either fullband or frequency-dependent TDOA, based on fixed smoothed statistics of the mixed DS and IS sources. The BM amplitude was determined optimally from adaptively smoothed statistics of the mixed sources, given the BM phase. The adaptive smoothing was provided as "Oracle" information in order to prevent the post-processing of the raw TDOA from influencing the results. Quantitative results measured over 30 seconds, comparing fullband and frequency-dependent SNE-PHAT, are shown in Table 3. Clearly, the frequency-dependent TDOA is relevant and offers better suppression of the DS. A qualitative illustration with the DS mixed with diffuse noise at 20 dB SNR is provided in Figure 4. The better suppression of the DS with frequency-dependent TDOA is evident from both the time domain waveform and spectrogram of the BM outputs.

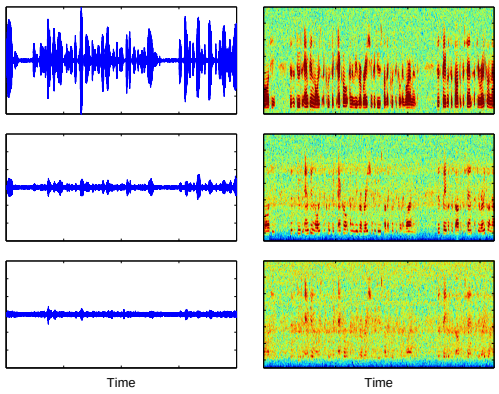

Fig. 4. Top to bottom: waveform (left) and spectrogram (right) of; a. BM input, BM output with, b. fullband TDOA, c. frequency-dependent TDOA.

The final experiment explores the tracking capability of a complete multi-microphone acoustic scene-analysis module using raw fullband TDOA estimates from SNE-PHAT. Acoustic signals were captured using the device in Figure 3: A primary talker (DS) was walking around a table with the device laying stationary on top while music (IS) was playing on loudspeakers. Figure 5 shows the scatter plot of raw TDOA estimates in black, and the trace of the TDOA of the inferred DS using acoustic scene analysis from [8] in green. In a complex acoustic scene, raw TDOA estimates capture direct path DS, DS reflections, and any IS (music in this case). Although the raw TDOA estimates appear scattered, they are clustered around the DS, and the cluster moves as the source moves at approximately 14 seconds. The DS TDOA trace in Figure 5 suggests that the post-processing method is able to correctly track the DS, based on the raw TDOA estimates from the SNE-PHAT method.

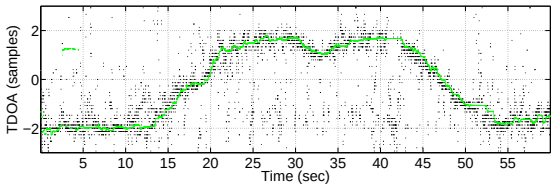

Fig. 5. Estimated TDOA values based on the raw TDOA values.

5. CONCLUSION

This paper presented a novel TDOA estimation algorithm inspired by null-steering. Experimental results with simulated data demonstrated that the SNE-PHAT method produces smaller localization errors relative to the GCC-PHAT method, particularly in the presence of noise. In addition, a generalization of fullband TDOA estimation to frequency-dependent TDOA estimation was presented, and results illustrated the benefits of using such frequency-dependent TDOA. The computational complexity discussion in the paper showed the proposed methods to be suitable for real-time implementation. In a final experiment with the SNE-PHAT method as part of a complete multi-microphone system with acoustic scene analysis to post-process raw TDOA data and infer desired and interfering sources, SNE-PHAT proved to result in good tracking of desired sources.

6. REFERENCES

[1] M. Brandestein and Darren Ward, *Microphone Arrays: Signal Processing Techniques and Applications (Digital Signal Processing)*, Springer, 2001.

[2] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transaction on Antennas Propagation*, vol. 30(1), pp. 27–34, 1982.

[3] I. Cohen, "Analysis of two-channel generalized sidelobe canceller (GSC) with post-filtering," *IEEE Transaction on Speech and Audio Processing*, vol. 11(6), pp. 684–699, 2003.

[4] C. Yemdji, *Acoustic echo cancellation for single-and dual-microphone devices: Application to mobile devices*, Ph.D. thesis, Thesis, June 2013.

[5] Y. Oualil, F. Faubel, M.M. Doss, and D. Klakow, "A TDOA gaussian mixture model for improving acoustic source tracking," *Proceedings of EUSIPCO*, pp. 1339–1343, 2012.

[6] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92(8), pp. 1950–1960, August, 2012.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55(3), pp. 216 – 228, 2007.

[8] J. Thyssen, A. Pandey, B.J. Borgstrom, D. Giacobello, and J-H Chen, "Multi-microphone source tracking and noise suppression," *US Patent and Trademark Office Publication Number 20140286497*, filed March 17, 2014, published September 25, 2014.

[9] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic scene localization," *Journal of Acoustical Society of America*, vol. 107, pp. 384 – 391, 2000.

[10] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transaction on Acoustics, Speech ,and Signal Processing*, vol. 24(4), pp. 320–327, 1976.

[11] T. Gustafsson, B.D. Rao, and M. Trivedi, "Source localization in reverberant environments: Performance bounds and ml estimation," in *Signals, Systems and Computers, 2001. Conference Record of the Thirty-Fifth Asilomar Conference on*, November 2001, vol. 2, pp. 1583–1587 vol.2.

[12] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, J. Neyman, Ed., pp. 481–492. University of California Press, Berkeley, CA, 1951.

[13] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *Journal of Acoustical Society of America*, vol. 122(6), pp. 3464 –3470, 2007.