SPEAKER TRACKING WITH SPHERICAL MICROPHONE ARRAYS

John McDonough[1], *Kenichi Kumatani*[1,2], *Takayuki Arakawa*[3], *Kazumasa Yamamoto*[4], *Bhiksha Raj*[1]

[1]Carnegie Mellon University, Pittsburgh, PA, USA
[2]Spansion, Inc., Sunnyvale, CA, USA
[3]NEC Corporation, Kawasaki-shi, Japan
[4]Toyohashi University of Technology, Toyohashi-shi, Japan

ABSTRACT

In prior work, we investigated the application of a spherical microphone array to a distant speech recognition task. In that work, the relative positions of a fixed loud speaker and the spherical array required for beamforming were measured with an optical tracking device. In the present work, we investigate how these relative positions can be determined automatically for real, human speakers based solely on acoustic evidence. We first derive an expression for the complex pressure field of a plane wave scattering from a rigid sphere. We then use this theoretical field as the predicted observation in an extended Kalman filter whose state is the speaker's current position, the direction of arrival of the plane wave. By minimizing the squared-error between the predicted pressure field and that actually recorded, we are able to infer the position of the speaker.

Index Terms— Microphone arrays, speech recognition, Kalman filters, spherical harmonics

1. INTRODUCTION

The state-of-the-art theory of beamforming with spherical microphone arrays explicitly takes into account two phenomena of sound propagation, namely, *diffraction* and *scattering*; see [1, §2] and [2, §6.10]. While these phenomena are present in all acoustic array processing applications, no particular attempt is typically made to incorporate them into conventional beamforming algorithms; rather, they are simply assumed to contribute to the room impulse response.

In prior work [3, 4, 5], we investigated the application of a spherical microphone array, the 32-channel Eigenmike®, to a distant speech recognition task. In that work, the relative positions of a fixed loud speaker and the spherical array required for beamforming were measured with an optical tracking device. In the present work, we investigate how these relative positions can be determined automatically for real human speakers based solely on acoustic evidence. For conventional microphone arrays, speaker tracking is typically performed by estimating time delays of arrival (TDOAs) between pairs of microphones using the phase transform [6] or adaptive eigenvalue decomposition [7]; the TDOAs can then be used as observations for a Kalman filter whose state corresponds to the speaker's position [8]. This approach works well for conventional arrays of modest dimensions because the signals arriving at any pair of microphones are—to a first approximation—time-shifted versions of another, which is equivalent to a phase shift in the frequency or subband domain. As we will discover in Section 2, such an approach is not suitable for rigid spherical arrays inasmuch the acoustics of such arrays introduce more complicated transformations of the signals arriving at pairs of sensors [9].

Meyer and Elko [9] were among the first authors to propose the use of spherical microphone arrays for beamforming. Initial work in source localization with spherical arrays used beamforming techniques to determine the three-dimensional power spectrum in a room and then applied peak search techniques to locate the dominant sources [9, 10]. Teutsch and Kellermann [11, 12] proposed to use eigenbeam ESPIRIT to perform source localization with cylindrical and spherical arrays; their approach was extended in [13] and more recently in [14].

In this work, we seek to develop an algorithm for speaker *tracking* as opposed to simple *localization*; this implies we will incorporate both past and present acoustic observations into the estimate of the speaker's current position as opposed to using merely the most recent observation. This is done to obtain a robust and smooth estimate of the speaker's time trajectory. To accomplish this objective, we first derive an expression for the complex pressure field of a plane wave scattering from a rigid spherical surface [2, §6.10.3]; this expansion is an infinite series of spherical harmonics appropriately weighted by the modal coefficients for scattering from a rigid sphere. We then use this theoretical field as the predicted observation in an extended Kalman filter whose state is the speaker's current position, which corresponds to the direction of arrival of the plane wave. By minimizing the squared-error between the predicted pressure field and that actually recorded at the sensors of the array, we are able to infer the position of the speaker. The Kalman filter provides for robust position estimates in that past observations are efficiently combined with the most recent one during the recursive correction stage.

We applied the proposed tracking algorithm to speech data spoken by real human speakers standing in front of a spherical microphone array. As the true speakers' positions are unknown, we evaluated the algorithm's effectiveness by performing beamforming using the estimated positions, then automatic speech recognition on the output of the beamformer. We found that our technique was able to reduce the final word error rate of the system from 50.9% using a single channel of the spherical microphone array to 45.4% using the beamformed array output for speech recognition.

The balance of this contribution is organized as follows. Section 2 reviews the derivation of an expression for the complex pressure field of a plane wave impinging on a rigid sphere; the final expression will involve an infinite series of spherical harmonics. Section 3 presents a speaker tracking system based on an extended Kalman filter that estimates the speaker's position by matching the actual, observed sound field impinging on a spherical array with that predicted by the theory of the preceding section. Empirical results are presented in Section 4 demonstrating the effectiveness of the proposed algorithm; the position estimates obtained with the

proposed algorithm are used for beamforming, and thereafter the enhanced speech signal from the beamformer is used for automatic recognition. In the final section, we briefly discuss the conclusions drawn from this work and our plans for future work.

2. ANALYSIS OF A PLANE WAVE IMPINGING ON A RIGID SPHERE

In this section, we develop a theoretical expression for the complex pressure field of a plane wave impinging on a rigid, spherical surface. We will also develop expressions for the partial derivative of this field with respect to the direction of arrival $\Omega = (\theta, \phi)$, where θ and ϕ denote the *polar angle* and *azimuth*, respectively. Let us express a plane wave impinging with a polar angle of θ on an array of microphones as [2, §6.10.1]

$$G_{\text{pw}}(kr, \theta, t) = e^{i(\omega t + kr\cos\theta)} = \sum_{n=0}^{\infty} i^n \, (2n+1) \, j_n(kr) \, P_n(\cos\theta) \, e^{i\omega t}, \quad (1)$$

where j_n and P_n are respectively the *spherical Bessel function* of the first kind and the *Legendre polynomial*, both of order n, $k \triangleq 2\pi/\lambda$ is the wavenumber, and $i \triangleq \sqrt{-1}$. Fisher and Rafaely [15] provide a similar expansion of spherical waves, such as would be required for near-field analysis. If the plane wave encounters a rigid sphere with a radius of a it is *scattered* [2, §6.10.3] to produce a wave with the pressure field

$$G_{\text{s}}(kr,ka,\theta,t) = - \sum_{n=0}^{\infty} i^n \, (2n+1) \, \frac{j_n'(ka)}{h_n'(ka)} \, h_n(kr) \, P_n(\cos\theta) \, e^{i\omega t}, \quad (2)$$

where $h_n = h_n^{(1)}$ denotes the *Hankel function* [16, §10.47] of the first kind while the prime indicates the derivative of a function with respect to its argument. Combining (1) and (2) yields the total sound pressure field [2, §6.10.3]

$$G(kr, ka, \theta) = \sum_{n=0}^{\infty} i^n (2n+1) \, b_n(ka, kr) \, P_n(\cos\theta), \quad (3)$$

where the nth *modal coefficient* is defined as

$$b_n(ka, kr) \triangleq j_n(kr) - \frac{j_n'(ka)}{h_n'(ka)} h_n(kr). \quad (4)$$

Note that the time dependence of (3) through the term $e^{i\omega t}$ has been suppressed for convenience. Plots of $|b_n(ka, ka)|$ for $n = 0, \ldots, 8$ are shown in Figure 1.

Let us now define the *spherical harmonic* of order n and degree m as [17]

$$Y_n^m(\theta, \phi) \triangleq \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-m)!}{(n+m)!}} \, P_n^m(\cos\theta) \, e^{im\phi}, \quad (5)$$

where P_n^m is the *associated Legendre function* of order n and degree m [18, §14.3]. The spherical harmonics fulfill the same role in the decomposition of square-integrable functions defined on the surface of a sphere as that played by the complex exponential $e^{i\omega nt}$ for decomposition of periodic functions defined on the real line. Let

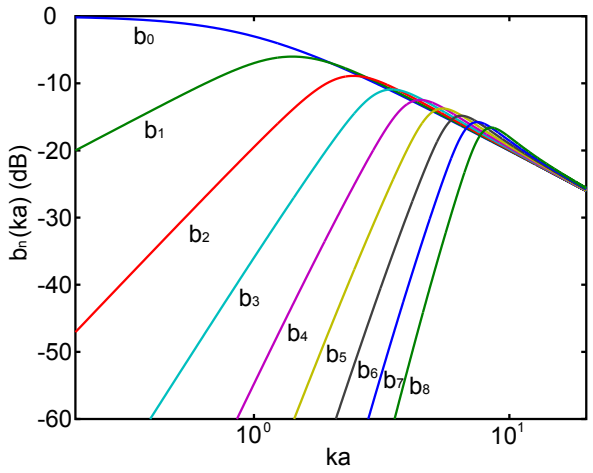

Fig. 1. Magnitudes of the modal coefficients $b_n(ka, ka)$ for $n = 0, 1, \ldots, 8$, where a is the radius of the sphere and k is the wavenumber.

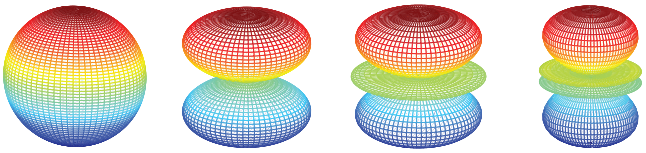

Fig. 2. The spherical harmonics Y_0, Y_1, Y_2 and Y_3.

γ represent the angle between the points (θ, ϕ) and (θ_s, ϕ_s) lying on a sphere, such that

$$\cos\gamma = \cos\theta_s \cos\theta + \sin\theta_s \sin\theta \cos(\phi_s - \phi). \quad (6)$$

Then the *addition theorem for spherical harmonics* [19, §12.8] can be expressed as

$$P_n(\cos\gamma) = \frac{4\pi}{2n+1} \sum_{m=-n}^{n} Y_n^m(\theta_s, \phi_s) \bar{Y}_n^m(\theta, \phi), \quad (7)$$

where $\bar{Y}$ denotes the complex conjugate of Y. Upon substituting (7) into (3), we find

$$G(kr_s,\theta_s,\phi_s,ka,\theta,\phi) = 4\pi \sum_{n=0}^{\infty} i^n \, b_n(ka, kr_s) \sum_{m=-n}^{n} Y_n^m(\theta_s, \phi_s) \bar{Y}_n^m(\theta, \phi), \quad (8)$$

where (θ, ϕ) denotes the direction of arrival of the plane wave and (r_s, θ_s, ϕ_s) denotes the position at which the sound field is measured. The spherical harmonics $Y_0 \triangleq Y_0^0$, $Y_1 \triangleq Y_1^0$, $Y_2 \triangleq Y_2^0$ and $Y_3 \triangleq Y_3^0$ are shown in Figure 2.

In all that follows, we will assume that $a = r_s$ so that ka and kr_s need not be shown as separate arguments. Based on the definition (5), we can write

$$\frac{\partial \bar{Y}_n^m(\theta, \phi)}{\partial \theta} = - \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-m)!}{(n+m)!}} \left. \frac{dP_n^m(x)}{dx} \right|_{x=\cos\theta} \cdot \sin\theta \cdot e^{-im\phi}, \quad (9)$$

$$\frac{\partial \bar{Y}_n^m(\theta, \phi)}{\partial \phi} = - \, i \, m \, \bar{Y}_n^m(\theta, \phi). \quad (10)$$

It remains to evaluate $dP_n^m(x)/dx$, which can be accomplished through the identity [18, §14.10]

$$(1-x^2)\frac{dP_n^m(x)}{dx} \equiv (m-n-1)P_{n+1}^m(x)+(n+1)xP_n^m(x). \quad (11)$$

These partial derivative expressions will be required for the linearization inherent in the *extended Kalman filter* (EKF).

3. SPEAKER TRACKING SYSTEM

Here we use development of the preceding section to formulate a complete tracking system based on the EKF. Let $\mathbf{y}_{k,l}$ denote a vector of *stacked* sensor outputs for the kth time step and the lth subband. Similary, let $\mathbf{g}_{k,l}(\theta,\phi)$ denote the model of the stacked sensor outputs

$$\mathbf{g}_{k,l}(\theta,\phi) \triangleq \begin{bmatrix} G(ka,\theta_0,\phi_0,ka,\theta,\phi) \\ G(ka,\theta_1,\phi_1,ka,\theta,\phi) \\ \vdots \\ G(ka,\theta_{S-1},\phi_{S-1},ka,\theta,\phi) \end{bmatrix}, \quad (12)$$

where $G(ka,\theta_s,\phi_s,ka,\theta,\phi)$ is given by (8). The linearization required to apply the EKF can then be expressed as

$$\frac{\partial G}{\partial \theta} = 4\pi \sum_{n=0}^{\infty} i^n b_n(ka) \sum_{m=-n}^{n} Y_n^m(\theta_s,\phi_s)\frac{\partial \bar{Y}_n^m(\theta,\phi)}{\partial \theta}$$

$$\frac{\partial G}{\partial \phi} = -4\pi \sum_{n=0}^{\infty} i^{n+1} b_n(ka) \sum_{m=-n}^{n} m Y_n^m(\theta_s,\phi_s)\bar{Y}_n^m(\theta,\phi).$$

The predicted observation inherent in the covariance form of the (extended) Kalman filter can then be formed from several components:

1. The individual sensor outputs given in (8); these are stacked as in (12).
2. A complex, time-varying frequency-dependent scale factor $B_{k,l}$, which is intended to model the unknown magnitude and phase variation of the subband components.
3. A complex exponential $e^{i\omega_l Dk}$, where ω_l is the center frequency of the lth subband and D is the decimation factor of the filter bank [20].

Given these definitions, the squared-error metric at time-step k can be expressed as

$$\epsilon(\theta,\phi,k) \triangleq \sum_{l=0}^{L-1} \left\| \mathbf{y}_{k,l} - \mathbf{g}_{k,l}(\theta,\phi)\, B_{k,l}\, e^{i\omega_l Dk} \right\|^2, \quad (13)$$

where $\mathbf{y}_{k,l}$ denotes the subband sensor outputs from a spherical array. Now note that if $B_{k,l}$ were known and (θ,ϕ) were treated as the state of a state-space system, then this time-varying state could be estimated with an extended Kalman filter; obviously the necessity of using an extended Kalman filter follows from the non-linearities in θ and ϕ evident in (5). It is readily shown that the maximum likelihood estimate of $B_{k,l}$ in (13) is given by

$$\hat{B}_{k,l} = \frac{\mathbf{g}_{k,l}^H(\theta,\phi)\mathbf{y}_{k,l}}{\left\|\mathbf{g}_{k,l}(\theta,\phi)\right\|^2} \cdot e^{-i\omega_l Dk}. \quad (14)$$

Note that if $\hat{B}_{k,l}$ in (14) is substituted into (13), the term $e^{i\omega_l Dk}$ will cancel out of the latter. Hence, these exponential terms can just as well be omitted from both (13) and (14).

Given the simplicity of (14), we might plausibly modify the standard extended Kalman filter as such:

1. Estimate the scale factors $B_{k,l}$ as in (14).
2. Use this estimate to update the state estimates $(\hat{\theta}_k,\hat{\phi}_k)$ of the Kalman filter.
3. (Possibly) perform an iterative update for each time step as in the *iterated extended Kalman filter* (IEKF) [21, §4.3.3] by repeating Steps 1 and 2.

We now briefly summarize the operation of the EKF. Let us state the *state* and *observation equations*, respectively, as

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{u}_{k-1}, \quad (15)$$

$$\mathbf{y}_k = \mathbf{H}_k(\mathbf{x}_k) + \mathbf{v}_k, \quad (16)$$

where $\mathbf{H}_k$ is the known, nonlinear *observation functional*. The noise terms $\mathbf{u}_k$ and $\mathbf{v}_k$ in (15–16) are by assumption zero mean, white Gaussian random vector processes with covariance matrices

$$\mathbf{U}_k = \mathcal{E}\{\mathbf{u}_k\mathbf{u}_k^H\}, \qquad \mathbf{V}_k = \mathcal{E}\{\mathbf{v}_k\mathbf{v}_k^H\},$$

respectively. Moreover, by assumption $\mathbf{u}_k$ and $\mathbf{v}_k$ are statistically independent. Let $\mathbf{y}_{1:k-1}$ denote all past observations up to time $k-1$, and let $\hat{\mathbf{y}}_{k|k-1}$ denote the minimum mean square error estimate of the next observation $\mathbf{y}_k$ given all prior observations, such that,

$$\hat{\mathbf{y}}_{k|k-1} = \mathcal{E}\{\mathbf{y}_k|\mathbf{y}_{1:k-1}\}.$$

By definition, the *innovation* is the difference $\mathbf{s}_k \triangleq \mathbf{y}_k - \hat{\mathbf{y}}_{k|k-1}$ between the actual and the predicted observations. This quantity is given the name innovation, because it contains all the "new information" required for sequentially updating the filtering density $p(\mathbf{x}_{0:k}|\mathbf{y}_{1:k-1})$ [21, §4]; i.e., the innovation contains that information about the time evolution of the system that cannot be predicted from the state space model.

We will now present the principal quantities and relations in the operation of the EKF; the details can be found in Haykin [22, §10], for example. Let us begin by stating how the predicted observation may be calculated based on the current state estimate, according to $\hat{\mathbf{y}}_{k|k-1} = \mathbf{H}_k(\hat{\mathbf{x}}_{k|k-1})$. Hence, we may write $\mathbf{s}_k = \mathbf{y}_k - \mathbf{H}_k(\hat{\mathbf{x}}_{k|k-1})$, which implies

$$\mathbf{s}_k = \bar{\mathbf{H}}_k(\hat{\mathbf{x}}_{k|k-1})\boldsymbol{\epsilon}_{k|k-1} + \mathbf{v}_k, \quad (17)$$

where $\boldsymbol{\epsilon}_{k|k-1} = \mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}$ is the *predicted state estimation error* at time k, using all data up to time $k-1$, and $\bar{\mathbf{H}}_k(\hat{\mathbf{x}}_{k|k-1})$ is the linearization of $\mathbf{H}_k(\mathbf{x})$ about $\mathbf{x} = \mathbf{x}_{k|k-1}$. It can be readily shown that $\boldsymbol{\epsilon}_{k|k-1}$ is orthogonal to $\mathbf{u}_k$ and $\mathbf{v}_k$ [22, §10.1]. Using (17) and exploiting the statistical independence of $\mathbf{u}_k$ and $\mathbf{v}_k$, the covariance matrix of the innovations sequence can be expressed as

$$\mathbf{S}_k \triangleq \mathcal{E}\left\{\mathbf{s}_k\mathbf{s}_k^H\right\} = \bar{\mathbf{H}}_k(\hat{\mathbf{x}}_{k|k-1})\mathbf{K}_{k|k-1}\bar{\mathbf{H}}_k(\hat{\mathbf{x}}_{k|k-1}) + \mathbf{V}_k, \quad (18)$$

where the *predicted state estimation error covariance matrix* is defined as

$$\mathbf{K}_{k|k-1} \triangleq \mathcal{E}\left\{\boldsymbol{\epsilon}_{k|k-1}\boldsymbol{\epsilon}_{k|k-1}^H\right\}. \quad (19)$$

The *Kalman gain* $\mathbf{G}_k$ can be calculated as

$$\mathbf{G}_k = \mathbf{K}_{k|k-1}\bar{\mathbf{H}}_k^H(\mathbf{x}_{k|k-1})\mathbf{S}_k^{-1}, \quad (20)$$

where the covariance matrix $\mathbf{S}_k$ of the innovations sequence is defined in (18). The *Riccati equation* then specifies how $\mathbf{K}_{k|k-1}$ can be sequentially updated, namely as,

$$\mathbf{K}_{k|k-1} = \mathbf{F}_{k|k-1}\,\mathbf{K}_{k-1}\,\mathbf{F}_{k|k-1}^H + \mathbf{U}_{k-1}. \quad (21)$$

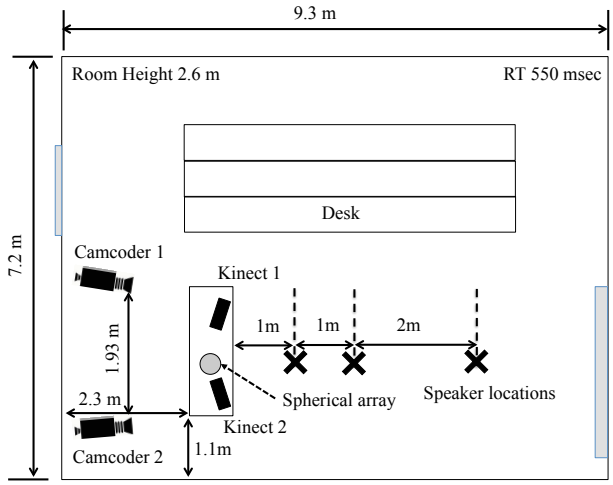

Fig. 3. Sensor configuration for data capture with the Eigenmike.

The matrix $\mathbf{K}_k$ in (21) is, in turn, obtained through the recursion,

$$\mathbf{K}_k = \mathbf{K}_{k|k-1} - \mathbf{G}_k\mathbf{H}_k\mathbf{K}_{k|k-1} = (\mathbf{I} - \mathbf{G}_k\mathbf{H}_k)\mathbf{K}_{k|k-1}. \quad (22)$$

This matrix $\mathbf{K}_k$ can be interpreted as the covariance matrix of the *filtered state estimation error* [22, §10], such that,

$$\mathbf{K}_k \triangleq \left\{\boldsymbol{\epsilon}_k\boldsymbol{\epsilon}_k^H\right\},$$

where $\boldsymbol{\epsilon}_k \triangleq \mathbf{x}_k - \hat{\mathbf{x}}_{k|k}$. Finally, filtered state estimate is given by

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{G}_k\mathbf{s}_k. \quad (23)$$

4. EXPERIMENTAL RESULTS

Figure 3 shows the sensor configuration used for our data capture. In the recording sessions, eleven human subjects are asked to read 25 sentences from the Wall Street Journal corpus at each of two different positions in order to investigate the sensitivity of recognition performance to the distance between speaker and array; as shown in the figure, the positions where the speaker was to stand were marked on the floor at 1 m, 2 m, and 4 m from the array measured parallel to the floor. The test data consisted of 6,948 words in total. The reverberation time T_{60} of the room was approximately 550ms. No noise was artificially added to the captured data, as natural noise from air conditioners, computers and other speakers was already present. The data was sampled at a rate of 44.1 kHz with a depth of three bytes per sample. Subband analysis was performed with the filter bank described in [21, §11] with $M = 512$ subbands.

The inter-sensor noise covariance matrix $\mathbf{V}_k$ for each subband required in (18) was estimated by analyzing segments of each session wherein the speaker was inactive with the filter bank [21, §11], summing the outer product of the subband snapshots, then scaling by the total number of frames analyzed. After the speakers positions were obtained with the speaker tracking system [8], beamforming was performed. We then ran the multi-pass speech recognizer described in [3] on the enhanced speech data. Table 1 shows *word error rates* (WERs) obtained with each beamforming algorithm as a function of distance between the speaker and the Eigenmike. As a reference, the word error rates obtained with the single array channel (SAC) and close-talking microphone (CTM) are also shown.

Algorithm	Distance	Pass (%WER)			
		1	2	3	4
SDM	1 m	75.6	43.6	31.6	28.8
	2 m	84.7	61.5	44.5	39.2
	4 m	89.4	72.5	57.1	50.9
SH SD BF	1 m	77.9	46.8	37.2	32.3
	2 m	83.2	58.2	43.3	39.0
	4 m	87.0	64.0	48.8	44.0
FF SD BF	1 m	79.7	50.9	38.3	35.6
	2 m	84.0	60.0	45.2	40.9
	4 m	85.5	67.4	49.8	45.4
CTM	Avg.	31.7	20.9	16.4	15.6

Table 1. WERs as a function of distances between the speakers and the Eigenmike.

Results are given in Table 1 for two variants of the super-directive beamformer. In the first, beamforming is performed in the spherical harmonics domain using the inter-harmonic covariance matrix derived by Yan et al. [23] for diffuse noise (SH SD BF). In the second variant, beamforming was performed directly on the sensor outputs without first performing modal analysis; moreover, the rigid spherical baffle was ignored inasmuch as the microphones were assumed to reside in a free field (FF SD BF). We found that spherical harmonics super-directive beamforming (SH SD BF) is more effective in terms of speech recognition performance.

The results reported in the table indicate that beamforming was ineffective at 1 m and 2 m, but provided a significant reduction in WER at 4 m. We attribute this to the far-field assumption used in derivinig (3); this assumption is largely valid for the distance of 4 m, but does not hold at the smaller distances. In future, we plan to investigate the near-field pressure field derived by Fisher and Rafaely [15] for speaker tracking and beamforming.

In the large vocabulary continuous speech recognition task, the distant speech recognizer with beamforming still lags behind the close talking microphone. However, the recognition performance can still be acceptable in applications that do not require recognizing every word precisely, such as dialogue systems.

5. CONCLUSIONS

Our results demonstrated that the combination of speaker tracking and beamforming enhanced the signal sufficiently to produce a significant reduction in the error rate of a distant speech recognition system when the speaker was located 4 m from the spherical array. For distances of 1 m and 2 m, however, a degradation in system performance was observed after tracking and beamforming. We attribute these contradictory results to the plane wave assumption we used in formulating our algorithm; such an assumption is valid for the greater distance, but not for the smaller. In future work, we will investigate the use of a near-field assumption for tracking and beamforming as in [15]. We also plan to compare our proposed method to other algorithms extant in the literature [10, 14, 24].

6. REFERENCES

[1] Heinrich Kutruff, *Room Acoustics*, Spoon Press, New York, NY, fifth edition, 2009.

[2] Earl G. Williams, *Fourier Acoustics*, Academic Press, San Diego, CA, USA, 1999.

[3] Kenichi Kumatani, John McDonough, and Bhiksha Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, pp. 127–140, November 2012.

[4] John McDonough and Kenichi Kumatani, "Microphone arrays," in *Techniques for Noise Robustness in Automatic Speech Recognition*, Tuomas Virtanen, Rita Singh, and Bhiksha Raj, Eds. Wiley, New York, NY, 2012.

[5] John McDonough, Kenichi Kumatani, and Bhiksha Raj, "Microphone arrays for distant speech recognition: Spherical arrays," in *Proc. APSIPA Conference*, Hollywood, CA, December 2012.

[6] G. C. Carter, "Time delay estimation for passive sonar signal processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 463–469, 1981.

[7] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Jour. of ASA*, vol. 107, no. 1, pp. 384–391, January 2000.

[8] U. Klee, G. Gehrig, and J.W. McDonough, "Kalman filters for time delay of arrival–based source localization," *Proc. of Eurospeech*, 2005.

[9] Jens Meyer and Gary W. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. ICASSP*, Orlando, FL, May 2002.

[10] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, "Spherical microphone array beamforming," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds. Springer, Berlin, 2010.

[11] Heinz Teutsch and Walter Kellermann, "Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2724–2736, 2006.

[12] Heinz Teutsch and Walter Kellermann, "Detection and localization of multiple wideband acoustic sources based on wavefield decomposition using spherical apertures," in *Proc. ICASSP*, Las Vegas, NV, USA, March 2008.

[13] Heinz Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*, Springer, Heidelberg, 2007.

[14] Haohai Sun, Heinz Teutsch, Edwin Mabande, and Walter Kellermann, "Robust localization of multiple sources in reverberant environments using EB-ESPIRIT with spherical microphone arrays," in *Proc. ICASSP*, Prague, Czech Republic, May 2011.

[15] Etan Fisher and Boaz Rafaely, "Near-field spherical microphone array processing with radial filtering," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 256–265, November 2011.

[16] Frank W. J. Olver and L. C. Maximon, "Bessel functions," in *NIST Handbook of Mathematical Functions*, Frank W. J. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark, Eds. Cambridge University Press, New York, NY, 2010.

[17] James R. Driscoll and Jr. Dennis M. Healy, "Computing Fourier transforms and convolutions on the 2-sphere," *Advances in Applied Mathematics*, vol. 15, pp. 202–250, 1994.

[18] T. M. Dunster, "Legendre and related functions," in *NIST Handbook of Mathematical Functions*, Frank W. J. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark, Eds. Cambridge University Press, New York, NY, 2010.

[19] George B. Arfken and Hans J. Weber, *Mathematical Methods for Physicists*, Elsevier, Boston, sixth edition, 2005.

[20] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, Englewood Cliffs, 1993.

[21] Matthias Wölfel and John McDonough, *Distant Speech Recognition*, Wiley, London, 2009.

[22] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, New York, fourth edition, 2002.

[23] Shefeng Yan, Haohai Sun, U. Peter Svensson, Xiaochuan Ma, and J. M. Hovem, "Optimal modal beamforming for spherical microphone arrays," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 2, pp. 361–371, 2011.

[24] D. Khaykin and B. Rafaely, "Coherent signals direction-of-arrival estimation using a spherical microphone array: Frequency smoothing approach," in *Proc. WASPAA*, New Paltz, NY, USA, October 2009.