TRACKING MULTIPLE TALKERS USING MICROPHONE-ARRAY MEASUREMENTS

Douglas E. Sturim[1] *Michael S. Brandstein*[2] *Harvey F. Silverman*[1] *

[1] Laboratory for Engineering Man/Machine Systems
Division of Engineering
Brown University
Providence, RI 02912
[2] Division of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138

ABSTRACT

A method for tracking the positional estimates of multiple talkers in the operating region of an acoustic microphone array is presented. Initial talker location estimates are provided by a time-delay-based localization algorithm. These raw estimates are spatially smoothed by a Kalman filter derived from a set of potential source motion models. Data association techniques based on the estimate clusterings and source trajectories are incorporated to match location observations with individual talkers. Experimental results are presented for array recorded data using multiple talkers in a variety of scenarios.

1. INTRODUCTION

The ultimate goal of this research is to passively track a number of talkers without the need for human operator control. The desired system should be capable of providing high-quality audio and visual data as sources move within a designated enclosure.

While a number of environmental cues are available for localizing, identifying, and tracking the sources of interest, this paper examines the use of acoustic information only for these applications. Arrays of microphones have been shown to be capable of accurately locating speech sources in a number of scenarios [1, 2, 3, 4, 5] while requiring significantly fewer computational resources in comparison to image-based systems. Furthermore, the microphone array can also be used to enhance the audio signals through the use of near-field beamforming.

The tracking methods detailed here take as their input the positional hypotheses provided from the time-delay-based location estimator addressed in [1]. This localization algorithm provides single-source position estimates at regular time intervals. During periods of constant, single-source speech activity, the rate of positional estimates is quite high (10-30 estimates/sec) using the Brown Megamike II system for 16 microphones [6]. During intervals of silence or direct, multiple talker overlap, no location observations are provided. However, the short independent analysis window (20-30 ms) and detection criteria associated with the algorithm make it effective for localizing moving sources and for situations where multiple talkers are active. In the former case, the positional observations follow the source motion. In the latter, the observations jump from one talker to another many times a second as one source dominates.

The algorithms presented in this paper concentrate on improving the quality of the positional estimates given a sequence of noise-corrupted location estimates from a set of multiple and, possibly, moving talkers. In such a scenario, the tracker must perform two main functions. The first is spatial filtering of the noise-corrupted location estimates. By employing a simple Newtonian motion model, these positional measurements may be smoothed via an appropriate Kalman filter. The motion of The second tracker function involves assigning the observations to their appropriate sources. In conjunction with the Kalman filter, data-association techniques can be used to assign measurements to individual talkers. Since the Kalman filter estimates the position of the talker within a calculable degree of certainty, the estimate can be used to combine each new measurement with an existing talker's past measurements. These distinct functions will be addressed in Sections 2 and 3 while some experimental results will follow in Section 4.

2. SPATIAL FILTERING

2.1. Kalman Filtering

The motion of a specified talker is modeled by the state difference equation:

$$\xi(k+1) = \mathbf{F}\xi(k) + \mathbf{G}\nu(k) \qquad (1)$$

The state, $\xi(k)$, is a 6-element vector consisting of some source's 3-dimensional Cartesian position and velocity:

$$\xi(k) = [x(k)\ y(k)\ z(k)\ \dot{x}(k)\ \dot{y}(k)\ \dot{z}(k)]^T \qquad (2)$$

It is evaluated at discrete iterations k which represent the continuous source motion sampled at regular intervals T. In the experiments that follow, T varies over the range $20 - 150ms$. A 3-element process noise vector, $\nu(k) = [\nu_x(k)\ \nu_y(k)\ \nu_z(k)]^T$, is used to model the non-zero acceleration of the source motion. It is composed of uncorrelated, zero-mean random variables with equal variance q^2 (i.e. $E[\nu(k)\nu(k)^T] = \mathbf{Q} = q^2\mathbf{I}_3$.) The transition matrix, $\mathbf{F}$, and the gain matrix, $\mathbf{G}$, are defined by:

$$\mathbf{F} = \left[\begin{array}{c|c} \mathbf{I}_3 & T\mathbf{I}_3 \\ \hline \mathbf{0}_3 & \mathbf{I}_3 \end{array}\right] \quad \mathbf{G} = \left[\begin{array}{c} \frac{T^2}{2}\mathbf{I}_3 \\ \hline T\mathbf{I}_3 \end{array}\right] \qquad (3)$$

The 3-dimensional source observation at the k^{th} iteration, $\vartheta(k)$, is modeled as the true source position corrupted by measurement noise, $w(k)$:

$$\vartheta(k) = \mathbf{H}\xi(k) + w(k) \qquad (4)$$

The covariance of the measurement noise, $\mathbf{R}(k)$, is calculated as a function of the source location, sensor positions, and background noise conditions [7]. The measurement matrix is given by $\mathbf{H} = [\ \mathbf{I}_3 \mid \mathbf{0}_3\ ]$

*Partially supported by NSF Grant MIP-9314625

Copyright 1997 IEEE

Given these state and observation difference equations, (1) and (4), the corresponding Kalman filter difference equations are summarized below. Detailed derivations of these equations can be found in a number of sources; [8] is typical.

One cycle of the Kalman filter is executed as follows:

Step 1: Prediction Equations- Given k observations the predicted state covariance at the $k+1$ iteration is calculated from:

$$\mathbf{P}(k+1|k) = \mathbf{F}\mathbf{P}(k|k)\mathbf{F}' + \mathbf{Q} \tag{5}$$

while the predicted measurement covariance is found from:

$$\mathbf{S}(k+1) = \mathbf{H}\mathbf{P}(k+1|k)\mathbf{H}' + \mathbf{R}(k) \tag{6}$$

The updated state covariance for $(k+1)$th iteration is calculated with:

$$\mathbf{P}(k+1|k+1) = \mathbf{P}(k+1|k) - \mathbf{W}(k+1)\mathbf{S}(k+1)\mathbf{W}'(k+1) \tag{7}$$

where $\mathbf{W}(k+1)$ is the filter gain given by:

$$\mathbf{W}(k+1) = \mathbf{P}(k+1|k)\mathbf{H}'(k+1)[\mathbf{S}(k+1)]^{-1} \tag{8}$$

Step 2: Gain Equations- The state prediction is found from:

$$\hat{\xi}(k+1|k) = \mathbf{F}\hat{\xi}(k|k) + \mathbf{G}u(k) \tag{9}$$

and the measurement prediction is given by:

$$\hat{\vartheta}(k+1|k) = \mathbf{H}\hat{\xi}(k+1|k) \tag{10}$$

The innovation of the filters, the difference between the $k+1$ observation $\vartheta(k+1)$ and measurement prediction $\hat{\vartheta}(k+1|k)$, is:

$$\nu(k+1) = \vartheta(k+1) - \hat{\vartheta}(k+1|k) \tag{11}$$

which is used to calculate the state estimate at the k+1 iteration:

$$\hat{\xi}(k+1|k+1) = \hat{\xi}(k+1|k) + \mathbf{W}(k+1)\nu(k+1) \tag{12}$$

The Kalman-filter is run by successively repeating steps 1 and 2 at each time iteration. Initialization of the filter is accomplished by the method of two-point differentiating from the initial measurements.

2.2. Interacting Multiple Model

Talkers using microphone arrays in typical environments can be stationary or wander about. Thus a single model, e.g., one for a stationary talker, for the kalman filter would not be ideal for all situations. Rather, a multiple model system should be better [9], and thus two models are used, a static model and a one for constant velocity. Both models should be statistical consistent within their respective realms of motion.

The variance of the process noise, q^2, may be scaled to reflect the accuracy of the constant-velocity model. For situations involving static talkers or sources moving at near constant velocity, a small process-noise variance is appropriate. When the source motion is "jerky" or erratic, a large variance is required to allow the Kalman tracker to accurately follow the true source motion. A mismatch between the model acceleration variance and the true source acceleration can result in large disparities between the tracked location and the source's actual position. Unlike many tracking scenarios, the motion of a talker is subject to a wide variation in acceleration. In practice, choosing a time-invariant q^2 value poses a difficult empirical problem. First, liberal overestimation of q^2 will allow the tracker to follow changes in motion, but will hinder the filter's ability to smooth the data. A second possibility would be to employ an adaptive process variance term. However the variation in acceleration may be extremely brief and such an adaptation with a one-step Kalman filter would not be possible. The Interacting Motion Model (IMM) is to maintain multiple motion models and choose between them based upon the observed data.

The IMM algorithm operates with two or more Kalman filters running in parallel. Each of these filters is derived from the motion model in (1) with a specific process noise variance. The algorithm transitions between filters according to a Markov chain in an effort to match the observed positional observations to the most appropriate motion model. Details of the procedure are given in [9].

3. DATA ASSOCIATION

Given multiple, possibly moving, talkers in an environment, the goal of data association is to relate isolated positional observations to a specific source. This study does not make a final decision as to which talker is most important at any single time. Instead it is assumed that the end user can select which talker is of most interest. While individual sources are spatially filtered via the Kalman filtering methods addressed in the previous section, each location estimate must first be associated with a particular source prior to the application of any spatial filtering. This is accomplished through the use of acceptance regions [8]. The acceptance region accounts for the measurement noise variance of the target and the possible motion of the target. A Kalman filter following a talker is called a "track". Each track or potential track is then placed into a "track table" where it can be sorted based on historical significance. If the Kalman filter has been following a talker for a few iterations then it is deemed historically significant.

The elements of the basic data association algorithm are outlined below.

1. **Track initialization:** A measurement that cannot be associated with any other tracks or other measurements is called an *initiator*. After the initialization, an acceptance region is set up for the second scan/iteration. If the next measurement falls within this acceptance region, a potential track is set up. The two measurements are then considered associated. The size of the acceptance region is a design parameter. The two associated measurements are used to initialize a Kalman-filter. The states and covariance associated with this track are stored within a track table. If an initiator is not associated with a measurement on the subsequent iteration, the potential track is dropped.
2. **Track continue:** Track acceptance regions are set up based on the predicted position when a Kalman-filter is running. If a new measurement falls within this acceptance region, then it used by the Kalman-filter. If no measurement falls within this acceptance region, then the predicted position is used as the measurement for the Kalman-filter. The output state and covariances of the Kalman-filter are then restored in the track table.

Copyright 1997 IEEE

3. **Track drop:** If a track does not have any measurements that fall within its acceptance region for a number of iterations, then the track is dropped. We define this number, N_d, as the number of iterations until the track is dropped. N_d is a user designed parameter and is based on how long the designer wants the track to continue when no measurements are present. For these experiments the tracks were held for about 1 sec or 30-50 iterations.

4. EXPERIMENTS

Recordings were made of talkers in a $4m \times 7m$ conference room with a large table in the center, a carpeted floor, and an acoustically tiled ceiling at a height of $2.75m$. Two eight-microphone, sub-arrays were positioned at standing height, $2m$ apart along one end of the conference table. Acoustic data from the 16 microphones was sampled at 20 kHz and positional estimates were found using the localization scheme detailed in [1]. One should note that for each frame, the locator yields a potential source location, if multiple sources are talking only the strongest source location is returned by the locator-estimator.

The first experiment involved a single, moving talker. Figure 1 displays an overhead view of the raw positional estimates for a talker walking toward one of the sub-arrays and the Kalman filtered version of this same data. For these experiments, the Interacting Multiple Model algorithm incorporated three potential source motion models: a static talker ($q^2 = 0$), and low and moderate acceleration talkers. The results illustrates the ability of the multiple-model Kalman filter to effectively smooth the positional measurements.

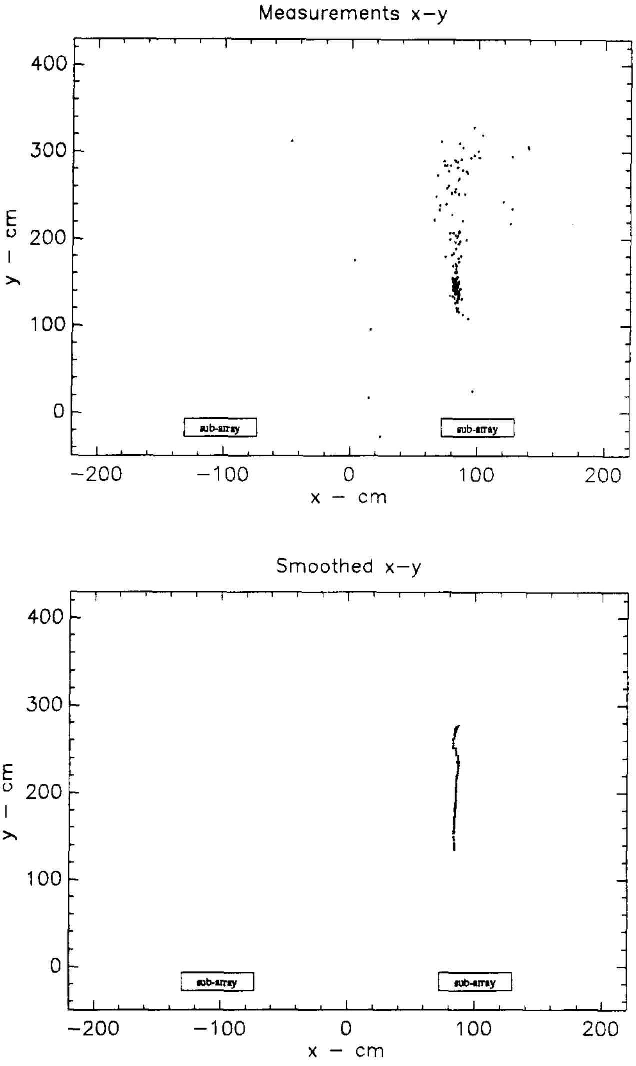
Figure 2 presents the data associating algorithm being run in a two-talker scenario. Here one talker is in motion while the other is stationary. The algorithm was able to assign each location estimate to a specific talker and then effectively smooth each track. Note the data association method's ability to reject erroneous measurements.

The final experiment considered four people sitting across from each other at a conference table. The conversation was a "round robin" of the speech spoken with some simultaneously. Talker one began with a brief 3-4 second statement followed by talker two continuing clockwise around the table, and so on. The entire conversation lasted 60 seconds. The raw positional observations and the results of running this data through the data association and IMM algorithm are presented in Figure 3

5. CONCLUSIONS

The Interacting Multiple model estimator is a viable technique for smoothing raw location estimates. The multiple-motion model offers a clear advantage as it effectively adapts to rapid changes in talker movement. This study was limited to only three motion models. Additional motion models as well as a more sophisticated source acceleration model may be straightforwardly incorporated incorporated into the IMM algorithm should performance considerations demand it. The data association methods described here were seen to be effective at distinguishing multiple talkers speaking simultaneously or in succession.

This work represents only the first step in constructing an automatic talker tracking system. One area of study is to extend the data association algorithm to include talker identification/characterization information. A statistical-modeling technique may be applied to the content of the source signal as in [10, 11]. Once the characterization of a talker has been developed, it will provide the data associator with additional information on which to make a decision. For example, a camera or the acoustic array could be steered to "the president".


Measurements x-y

Smoothed x-y

Figure 1. Overhead plots of the positional measurements for a single talker moving toward a sub-array before (top) and after (bottom) Kalman filtering.

The methods discussed here have focussed entirely on acoustic data. A hybrid-system which fused both sound and image data could have many potential advantages. For instance, using location estimates derived from the microphone signals to limit the search region of a visual-based tracking system could provide for significant improvements in functionality and computational efficiency. Also, the system would be effective when the talkers are quiet.

REFERENCES

[1] M. S. Brandstein. *A Framework for Speech Source Localization Using Senor Arrays*. PhD thesis, Brown University, Providence RI, 1995.

[2] M. S. Brandstein, J. E. Adcock, and H. F. Silverman. A practical time-delay estimator for localizing speech

Copyright 1997 IEEE

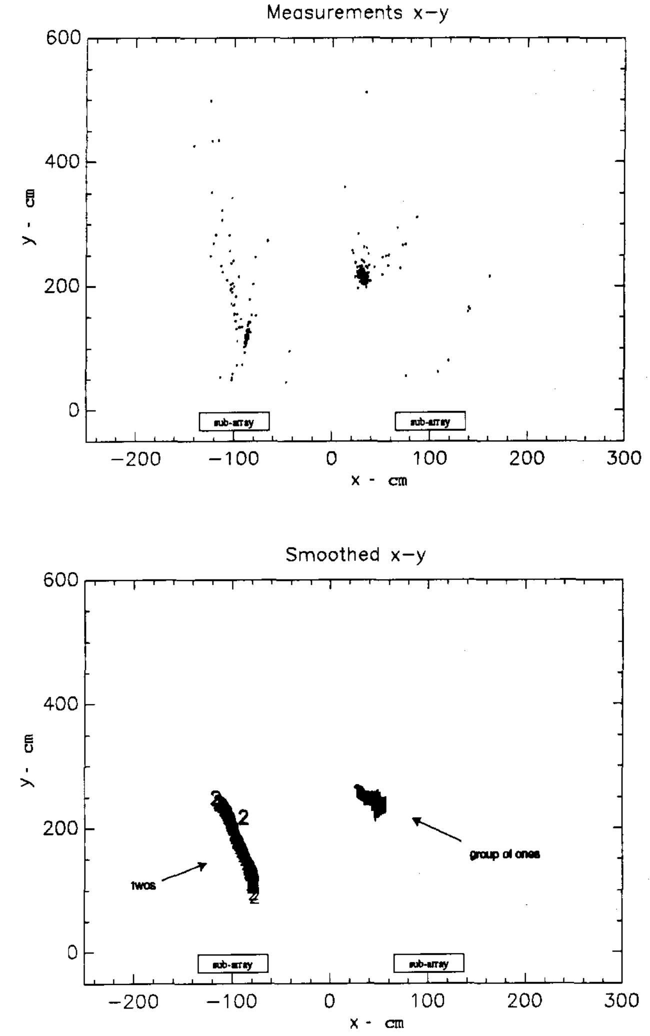


Figure 2. Overhead plots of the positional measurements for two talkers, one static and the other moving, before (top) and after (bottom) data association and Kalman filtering.

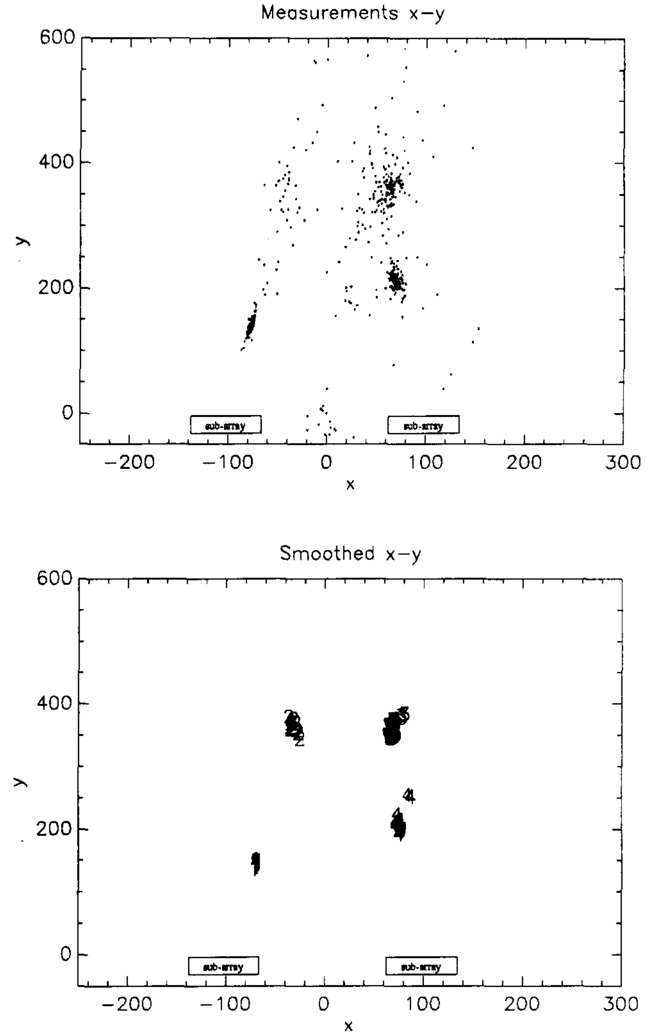


Figure 3. Overhead plots of the positional measurements for four talkers before (top) and after (bottom) data association and Kalman filtering.

sources with a microphone array. *Computer, Speech and Language*, 9(2):153–169, April 1995.

[3] M. Omologo and P. Svaizer. Acoustic source location in noisy and reverberant environment using csp analysis. In *Proceedings of ICASSP96*, pages 921–924. IEEE, 1996.

[4] H. F. Silverman and S. E. Kirtman. A two-stage algorithm for determining talker location from linear microphone-array data. *Computer, Speech, and Language*, 6(2):129–152, April 1992.

[5] J. Flanagan, D. Berkley, G. Elko, J. West, and M. Sondhi. Autodirective microphone systems. *Acustica*, 73:58–71, 1991.

[6] J. E. Kessler. On the implementation of a real-time microphone-array audio source location algorithm with applications to video teleconferencing. Master's thesis, Brown University, Providence, RI, May 1996.

[7] M. S. Brandstein, J. E. Adcock, and H. F. Silverman. Microphone array localization error estimation with application to optimal sensor placement. *J. Acoust. Soc. Am.*, Vol.99(6):3807–3816, 1996.

[8] Y. Bar-Shalom and T. E. Fortman. *Tracking and Data Association*. Academic Press, 1988.

[9] D. T. Lerro and Y. Bar-Shalom. Interacting multiple model tracking with target amplitude features. *IEEE Trans. AES*, Vol.29(2):494–509, 1993.

[10] D. A. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Trans. SAP*, 2(4):639–643, October 1994.

[11] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang. A vector quantization approach to speaker recognition. *Proceedings of ICASSP85*, pages 387–390, 1985.