SPEAKER TRACKING FOR TELECONFERENCING VIA BINAURAL HEADSET MICROPHONES

Hannes Gamper, Sakari Tervo and Tapio Lokki

Department of Media Technology
Aalto University School of Science, Espoo, Finland
first [dot] last [at] aalto.fi

ABSTRACT

Speaker tracking for teleconferencing via user-worn binaural headset microphones in combination with a reference microphone array is proposed. The tracking is implemented with particle filtering based on maximum likelihood estimation of time-difference of arrival (TDOA) estimates. An importance function for prior weighting of the particles of silent conferees (i.e., "listeners") is proposed. Experimental results from tracking three conferees in a meeting scenario are presented. The use of user-worn microphones in addition to a reference microphone array is shown to improve speaker distance estimation and overall tracking performance substantially. The importance function improved the tracking RMSE by 58% on average. The position tracking RMSE of the proposed method is about 0.11 m.

Index Terms— Speaker tracking, acoustic source tracking, teleconferencing, binaural headset microphones

1. INTRODUCTION

Information about the location and orientation of participants in a teleconference scenario is useful to allow spatial interaction between the conferees, or for tools to counteract the effect of reverberation [1]. Acoustic source tracking setups typically include several microphones, distributed across the room [2, 3] or arranged in clusters or arrays [4, 5]. This allows reliable tracking of the speakers, given that the acoustic conditions are favourable [2]. Tracking the position and head orientation of a user via binaural headset microphones was previously proposed using anchor sound sources at known positions [6]. For reliable speaker tracking, the aforementioned systems require complex and costly installation of multiple arrays, or anchor sources at known positions.

In this paper, we propose a tracking system for teleconferencing that relies on a single microphone array and user-worn microphones. The tracking is entirely based on the speech signals of the conferees, recorded at binaural headset microphones and a reference microphone array. A method for tracking the head orientation of the conferees for the given scenario was previously proposed by the present authors [7], hence this article concentrates on location tracking. The proposed method for tracking the speaker positions in a conference scenario consists of three parts. First, basic voice activity detection is performed to determine the active speaker from the binaural microphone signals. It relies on thresholding the signal energy recorded at the binaural headset microphones and the tracking evidence found for the particles of each conferee to determine who spoke and when. Then, particle filtering is applied to track the position of the active speaker. Finally, the distance of each conferee to the active speaker is estimated to derive an importance function for prior weighting of the particles of silent conferees.

The contribution of this article is twofold. Firstly, employing user-worn microphones for speaker tracking is proposed and the resulting tracking accuracy improvement is shown. Secondly, a prior weighting method of the particles of silent conferees (i.e., the listeners) is proposed. It is based on deriving an importance function from the distance of each listener to the active speaker estimated from the signals of user-worn headset microphones. In an experimental setup the locations of three conferees (two seated, one moving) engaged in a lively discussion were tracked. The root-mean-square error (RMSE) for the speaker tracking was about 0.11 m using two binaural headset microphones per conferee, and about 0.13 m using one binaural headset microphone per conferee. In both cases, the proposed importance function for prior particle weighting of the inactive conferees led to equal or improved tracking performance. Speaker tracking without user-worn microphones resulted in an RMSE of several metres, mainly due to speaker distance estimation errors, indicating a substantial improvement in tracking accuracy through the usage of user-worn microphones. The results show the proposed methods for speaker tracking and prior weighting of particles to be reasonably robust and accurate.

2. MOTIVATION AND PROPOSED METHOD

Acoustic source tracking in office environments is challenging due to noise and reverberation. Here we propose a tracking method combining a microphone array and user-worn microphones. The advantage of integrating user-worn microphones into the tracking system is their vicinity to the acoustic source, i.e., the speaker, which in turn can result in better signal-to-noise ratio and hence better raw data for the acoustic source tracking. Furthermore, the distance between the user-worn microphones and the speaker is relatively constant. Thus, the speaker–array and speaker–listener distances can be estimated from time-difference of arrival (TDOA) estimates. The tracking system proposed here takes advantage of distance estimates obtained from the headset microphones for improved tracking accuracy and robustness. The tracking is implemented via particle filtering, a technique that can be used to recursively approximate the probability density of an unknown source location [1]. The steps involved in acoustic source tracking via particle filtering are [1]: a) obtaining localisation measurements by applying a localisation function to the

The research leading to these results has received funding from the Academy of Finland, project nos. [218238 and 140786], the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no. [203636], the Helsinki Graduate School in Computer Science and Engineering (HECSE), the Nokia Research Foundation, and [MIDE program] of Aalto University.

microphone signals; b) forming a likelihood function based on the localisation measurements; c) weighting the particles of the particle filter according to the likelihood function. The source location estimate is obtained as the weighted sum of the particle locations.

2.1. Localisation measurements

For each pair of microphones, an estimate of the time-difference of arrival (TDOA) $\hat{\tau}$ of the source signal is determined. Here, the TDOA estimates are obtained via the generalised correlation framework with phase transform [8].

2.2. Geometric relations

To calculate the likelihood of observing the estimated TDOAs $\hat{\tau}$ with a speaker at position $\mathbf{s}$, the expected TDOAs $\tau(\mathbf{s})$ need to be calculated. The TDOA τ of microphones i and j is calculated as the difference of two time of arrivals (TOAs) t of the speech signal:

$$\tau_{i,j}(\mathbf{s}) = t_i(\mathbf{s}) - t_j(\mathbf{s}) = c^{-1}\|\mathbf{s} - \mathbf{m}_i\| - c^{-1}\|\mathbf{s} - \mathbf{m}_j\|, \quad (1)$$

where $\mathbf{s}$ and $\mathbf{m}$ denote the speaker and microphone positions, respectively. The TDOA τ between a microphone i of the reference microphone array and the left and right binaural headset microphones $h_{\mathrm{l,sp}}$ and $h_{\mathrm{r,sp}}$ worn by the speaker is given as

$$\begin{aligned}\tau_{i,h_{\mathrm{l,sp}}}(\mathbf{s}) &= c^{-1}\|\mathbf{s} - \mathbf{m}_i\| - c^{-1}\|\mathbf{s} - \mathbf{m}_{h_{\mathrm{l,sp}}}\| \\ \tau_{i,h_{\mathrm{r,sp}}}(\mathbf{s}) &= c^{-1}\|\mathbf{s} - \mathbf{m}_i\| - c^{-1}\|\mathbf{s} - \mathbf{m}_{h_{\mathrm{r,sp}}}\| \end{aligned} \quad (2)$$

where $c = 343$ m/s is the speed of sound. We assume the distance from the speaker to his or her binaural microphones to be fixed and estimate it as $d_{\mathrm{s}} = \|\mathbf{s} - \mathbf{m}_{h_{\mathrm{l,sp}}}\| = \|\mathbf{s} - \mathbf{m}_{h_{\mathrm{r,sp}}}\| \approx 0.18$ m, which corresponds to 0.5 ms delay. Assuming equal distance of the speaker to both headset microphones, i.e., $\tau_{h_{\mathrm{l,sp}},h_{\mathrm{r,sp}}} = 0$, the time of arrival (TOA) t from the speaker $\mathbf{s}$ to the microphone i is given as

$$t_i(\mathbf{s}) = c^{-1}d_{\mathrm{s}} + \tau_{i,h_{\mathrm{l,sp}}}(\mathbf{s}) = c^{-1}d_{\mathrm{s}} + \tau_{i,h_{\mathrm{r,sp}}}(\mathbf{s}). \quad (3)$$

2.3. Likelihood function

The likelihood function for a microphone pair i, j, speaker position $\mathbf{s}$, and expected and estimated TDOA $\tau_{i,j}$ and $\hat{\tau}_{i,j}$ is given as [1]

$$p(\tau_{i,j}(\mathbf{s})|\hat{\tau}_{i,j}, \sigma_{i,j}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\tau_{i,j}(\mathbf{s}) - \hat{\tau}_{i,j})^2}{2\sigma_{i,j}^2}\right), \quad (4)$$

i.e., a normal distribution with variance $\sigma_{i,j}^2$ and mean $\hat{\tau}_{i,j}$. The maximum likelihood estimation (MLE) function is given as [1]

$$p(\mathbf{s}) = \prod_{\{i,j\}=1}^{M} p(\tau_{i,j}(\mathbf{s})|\hat{\tau}_{i,j}, \sigma_{i,j}), \quad (5)$$

where M denotes the number of microphone pairs. With (3), the MLE function for a combination of binaural headset microphones worn by the speaker $h_{\mathrm{l,sp}}$ and $h_{\mathrm{r,sp}}$ and N reference array microphones is given as

$$\begin{aligned} p(\mathbf{s}) = &\prod_{i=1}^{N}\prod_{j=1}^{2} p(t_i(\mathbf{s})|c^{-1}d_{\mathrm{s}} + \hat{\tau}_{i,h_{j,\mathrm{sp}}}, \sigma_{i,h_{j,\mathrm{sp}}})\times \\ &\prod_{\{i,j\}=1}^{M} p(\tau_{i,j}(\mathbf{s})|\hat{\tau}_{i,j}, \sigma_{i,j}). \end{aligned} \quad (6)$$

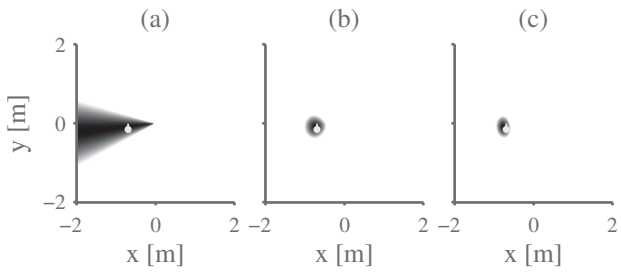

Fig. 1. MLE function for speaker tracking obtained from a reference microphone array at $(0, 0)$ and (a) no, (b) one, and (c) two user-worn headset microphones. The light-grey dot indicates the true position and head orientation of the speaker.

An example of the MLE function in the tracking area with no, one, and two binaural headset microphones is shown in Figure 1. There is a bias in the speaker location estimation, as the speaker's acoustic centre does not necessarily coincide with the speaker's position. The bias is visible in Figure 1 as a shift in the likelihood towards the front of the speaker.

2.4. Particle filtering

The particle filter is initialised by forming a set of K particles uniformly distributed in the tracking area. The filtering consists of three steps: prediction, update, and resampling. To predict the movement of the particles, Brownian motion is assumed in this article, hence the particles are propagated according to a random distribution [2]. In the update step, a weight w is calculated for each particle k at location $\mathbf{l}$ with (6) as $w_k = p(\mathbf{l}_k)$. The weights are normalised so that $\sum_k^K w_k = 1$. The speaker location estimate $\hat{\mathbf{s}}$ is given as the weighted sum of the particle locations $\mathbf{l}$:

$$\hat{\mathbf{s}} = \sum_{k=1}^{K} w_k \mathbf{l}_k. \quad (7)$$

The resampling of the particles is done with stratified resampling according to the weights [9].

2.5. Importance function for silent conferees ("listeners")

The TOA from the speaker $\mathbf{s}$ to the listener's left binaural headset microphone $h_{\mathrm{l,lis}}$ is given by:

$$t_{h_{\mathrm{l,lis}}}(\mathbf{s}) = c^{-1}d_{\mathrm{s}} + \tau_{h_{\mathrm{l,lis}},h_{\mathrm{l,sp}}}(\mathbf{s}) = c^{-1}d_{\mathrm{s}} + \tau_{h_{\mathrm{l,lis}},h_{\mathrm{r,sp}}}(\mathbf{s}),$$

where $h_{\mathrm{l,sp}}$ and $h_{\mathrm{r,sp}}$ denote the speaker's left and right headset microphones. The calculation of $t_{h_{\mathrm{r,lis}}}$ is analogous. The TDOAs between the listener headset microphones and a reference array microphone i, i.e., $\tau_{i,h_{l,lis}}$ and $\tau_{i,h_{r,lis}}$, can be calculated using (1,2,3). From these geometric relations, an importance function p_{I} for the listener particles can be derived as

$$\begin{aligned} p_{\mathrm{I}}(\mathbf{s}) = &\prod_{i=1}^{2}\prod_{j=1}^{2} p(t_{h_{i,\mathrm{lis}}}(\mathbf{s})|c^{-1}d_{\mathrm{s}} + \hat{\tau}_{h_{i,\mathrm{lis}},h_{j,\mathrm{sp}}}, \sigma_{h_{i,\mathrm{lis}},h_{j,\mathrm{sp}}})\times \\ &\prod_{i=1}^{N}\prod_{j=1}^{2} p(\tau_{i,h_{j,\mathrm{lis}}}(\mathbf{s})|\hat{\tau}_{i,h_{j,\mathrm{lis}}}, \sigma_{i,h_{j,\mathrm{lis}}}) \end{aligned} \quad (8)$$

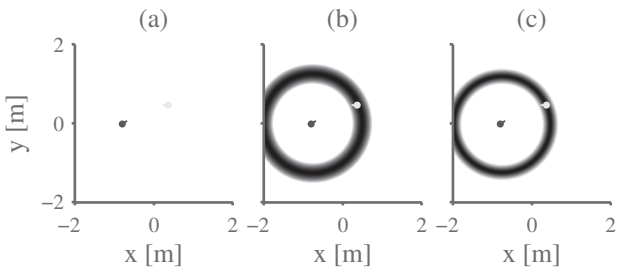

Fig. 2. Importance function for listener P2 with (a) no, (b) one, and (c) two user-worn microphones. It is derived from the estimated distance to the active speaker P1 and used for prior weighting of the particles of P2. The dark-grey and the light-grey dot indicate the true position and head orientation of P1 and P2, respectively.

The importance function yields the listener particle weights, indicating where the listener particles are to be sampled [1]. Figure 2 illustrates an example of the MLE function for one listener with both speaker and listener wearing no, one, and two binaural headset microphones. The importance function has the form of a circle, centered at the estimated speaker location, with a radius corresponding to the estimated speaker–listener distance. Using just one binaural headset microphone, the head orientation of the listener introduces a bias of max. ± 0.1 m (i.e., half the head radius) to the estimated speaker–listener distance (cf. Figure 2(b)).

3. EXPERIMENTAL SETUP

In a case study, the positions of three conferees were tracked during 60 seconds of a conversation in a meeting scenario (cf. Fig. 3). Each participant was wearing a binaural headset with integrated microphones of type Philips SHN2500. The reference microphone array located in the centre of the tracking area at $(0, 0)$ was of type G.R.A.S VI 50, consisting of six microphones, two on each axis with 100 mm spacing. The microphone signals were recorded at a sampling rate of 96 kHz. The recording was made in a multipurpose space with a reverberation time of about 0.3 s. The signal-to-noise ratio (SNR) was between 15 and 30 dB during active speech frames. The ground truth data for the position tracking was obtained from a video recorded via a Canon EOS 7D camera mounted about 4 m above the scene. Visually distinct markers placed on the head of each conferee were tracked in the video stream using the ARToolkit, which for the given setup provides a tracking accuracy of around 1 cm [10] at an update rate of 30 Hz. Each conferee was tracked using a particle filter with $K = 100$ particles. The standard deviations used in the MLE functions were $\sigma_{i,j} = 0.5$, $\sigma_{i,h_{j,\mathrm{sp}}} = \sigma_{h_{i,\mathrm{lis}},h_{j,\mathrm{sp}}} = \sigma_{i,h_{j,\mathrm{lis}}} = 5$.

4. RESULTS

The root-mean-square errors (RMSE) for the speaker tracking under various conditions are presented in Table 1. Best performance for the position tracking is achieved using two binaural headset microphones for each conferee and importance functions for the particles of silent conferees (i.e., the listeners). The tracking performance for this condition is shown in Figure 4. With each conferee wearing just one microphone the performance deteriorates slightly. The use of the importance function has a larger effect on the tracking per-

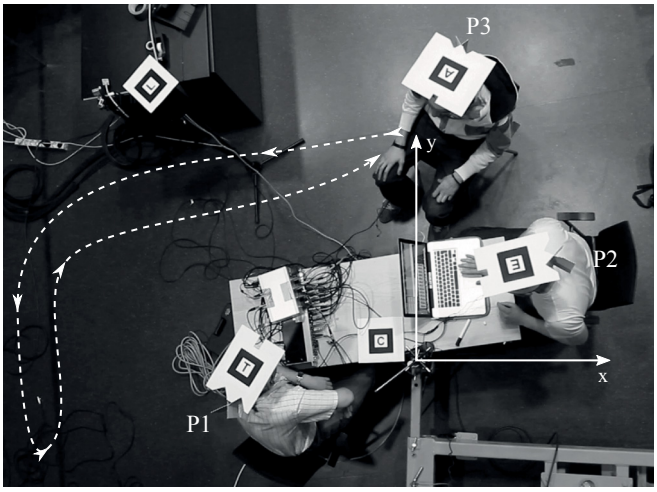

Fig. 3. The experimental setup of the case study. The dashed line illustrates the path of conferee P3. The reference microphone array is located at the centre of the coordinate axes.

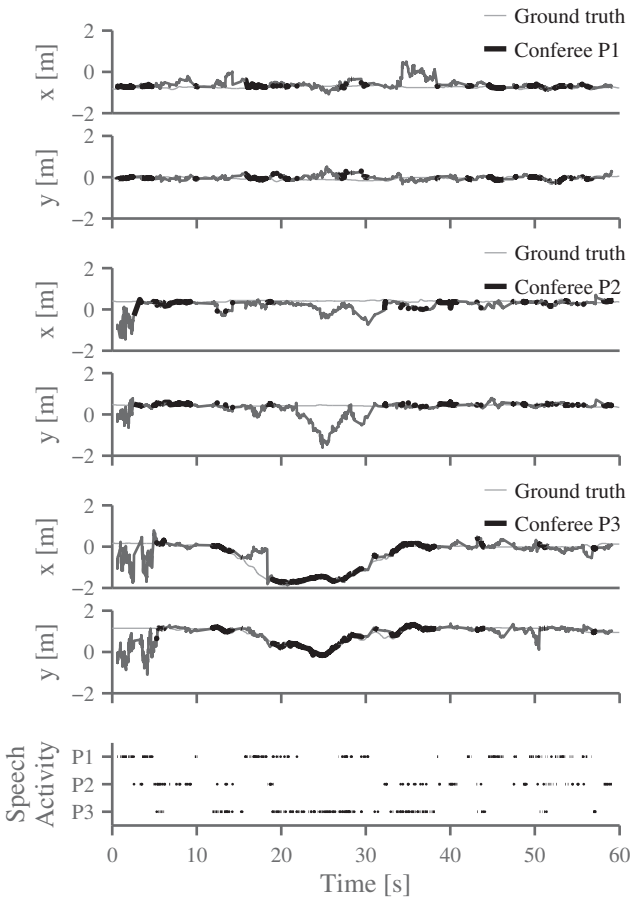

Fig. 4. Tracking results and speech activity map. The active speaker, marked in the activity map with a dot for each frame, was detected as the conferee with the maximum headset microphone energy and sum of non-normalised particle weights. Speech activity was detected in 38% of all frames, with reasonable accuracy. The tracking results during frames where a conferee was active are marked as bold lines.

Table 1. Tracking performance. M denotes the number of binaural headset microphones used for each conferee.

	Distance RMSE [m]					Angle RMSE [deg]					Position RMSE [m]				
M	0	1	1*	2	2*	0	1	1*	2	2*	0	1	1*	2	2*
P1	5.53	0.07	0.11	**0.07**	0.08	8.04	10.74	9.53	8.46	**6.88**	5.54	0.15	0.15	0.13	**0.11**
P2	10.81	0.07	0.06	**0.05**	0.05	26.18	23.25	12.89	22.25	**11.54**	10.87	0.22	0.13	0.21	**0.12**
P3	1.61	**0.06**	0.09	0.09	0.07	10.42	15.69	**3.69**	7.06	3.73	1.64	0.39	0.12	0.14	**0.11**
mean	5.99	**0.07**	0.09	0.07	0.07	14.88	16.56	8.70	12.59	**7.38**	6.02	0.26	0.13	0.16	**0.11**

*Prior weighting of listener particles based on importance function.

Table 2. Tracking performance compared to state of the art tracking systems tested under similar experimental conditions.

Method	SBF-PL [11]	SBF-TBD [12]	MI-BF[5]	**here**
RMSE [m]	0.14	0.29	0.14	**0.11**

formance for both the one- and two-microphone condition, reducing the position tracking RMSE for both conditions by 58% on average. Position tracking performance without user-worn binaural headset microphones is poor, due to the small spacing of the reference microphone array, which makes a robust localisation and speaker distance estimation challenging (cf. Table 1). With the use of binaural headset microphones, the distance RMSE is below 0.09 m on average for all conditions, i.e., in a similar range as the head radii of the conferees. Without importance function, the angle RMSE does not improve with the addition of binaural headset microphones. Using the importance function improves the angle RMSE clearly compared to the reference microphone array alone, indicating that the importance function improves speaker direction tracking. A succession of importance functions obtained from different speakers forces the particles of a listener to cumulate at the intersection points of the importance functions. One of the intersection points lies at the true location of the listener, thus allowing a rough estimate of the listener location from the particle locations (cf. Figure 4, 20–30 s: tracking for the silent P2 re-converges to the true location around 28 s).

5. CONCLUSION

This article studies the location tracking of speakers via binaural headset microphones and a reference microphone array. The tracking RMSE of the proposed system is comparable to values reported for state of the art tracking systems under similar experimental conditions (see Table 2). The tracking performance for active speakers is greatly improved through the addition of user-worn binaural headset microphones, mainly due to improved speaker distance estimation. The tracking performance is further improved through the use of importance functions for particles of silent conferees, derived from the headset microphone signals. For the given setup, the root-mean-square error (RMSE) of the speaker tracking is about 0.11 m for both static and moving conferees. Using just one user-worn headset microphone per conferee, the RMSE for speaker tracking is 0.13 m. This suggests that the proposed tracking framework might work with other forms of user-worn microphones, e.g. lavalier microphones attached to the clothing of the conferees. Future work includes testing the setup in noisier and/or more reverberant environments, and implementing improved speech activity detection.

6. REFERENCES

[1] E.A. Lehmann, *Particle filtering methods for acoustic source localisation and tracking*, Ph.D. thesis, Australian National University, 2004.

[2] P. Pertilä, T. Korhonen, and A. Visa, "Measurement combination for acoustic source localization in a room environment," *EURASIP J. Audio, Speech, and Music Process.*, vol. 2008, pp. 3, 2008.

[3] Kook Cho, T. Nishiura, and Y. Yamashita, "Robust speaker localization in a disturbance noise environment using a distributed microphone system," in *IEEE ISCSLP, Tainan, Taiwan*, 2010, pp. 209–213.

[4] H. Sun, S. Yan, and P. Svensson, "Robust spherical microphone array beamforming with multi-beam-multi-null steering, and sidelobe control," in *Proc. IEEE WASPAA-09, New Paltz, NY*, oct. 2009, pp. 113–116.

[5] F. Talantzis, "An acoustic source localization and tracking framework using particle filtering and information theory," *IEEE Trans. Audio,. Speech, Lang. Process.*, vol. 18, no. 7, pp. 1806–1817, sept. 2010.

[6] M. Tikander, A. Härmä, and M. Karjalainen, "Binaural positioning system for wearable augmented reality audio," in *Proc. IEEE WASPAA-03, New Paltz, NY*, 2003, pp. 153–156.

[7] H. Gamper, S. Tervo, and T. Lokki, "Head orientation tracking using binaural headset microphones," in *Proc. of the AES 131st Conv., New York, NY*, 2011.

[8] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[9] R. Douc, O. Cappé, and E. Moulines, "Comparison of resampling schemes for particle filtering," in *Proc. IEEE ISPA-05, Zagreb, Croatia*, 2005, pp. 64–69.

[10] H. Kato and M. Billinghurst, "Marker tracking and hmd calibration for a video-based augmented reality conferencing system," in *Proc. IEEE IWAR, San Francisco, CA*, 1999, pp. 85–94.

[11] D.B. Ward, E.A. Lehmann, and R.C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 826–836, 2003.

[12] M.F. Fallon and S. Godsill, "Acoustic source localization and tracking using track before detect," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1228–1242, 2010.