

Speech-based localization of multiple persons for an interface robot

Gradje Klaassen Wojciech Zajdel Ben J.A. Kröse
ISLA, Informatics Institute
University of Amsterdam
Kruislaan 403, 1098SJ Amsterdam, The Netherlands
{gklaasse,wzajdel,krose}@science.uva.nl

Abstract—Robots are conveniently controlled by a human operator with spoken commands, since voice is a natural communication medium for humans. In order to successfully carry out a command, a robot needs to know which of the possibly many people gave the command and where this person is located. In this paper we present a particle-filter based algorithm for localization of multiple speakers, in an environment where there is only one person speaking at a time. The algorithm incorporates person-specific voice features (vowel formant frequencies) in order to distinguish between the speakers. The voice features are supported by azimuth angle measurements obtained by a pair of microphones. We test our approach using the microphone system of the Philips iCat interface robot.

Index Terms—Multi-target tracking, Speaker localization, Bayesian filtering, Human-robot interaction.

I. INTRODUCTION

One of the aspects of human-robot interaction is to interpret in an intelligent manner human voice perceived by a robot. This problem becomes more important, as current robots are moving out of the factory floor into environments inhabited by human. Examples are museum or exhibition robots [1], [17], care-for-elderly robots [12], office [2] and entertainment robots [1]. In their role as guide, companion or servant these systems have to interact with the humans they encounter.

As a part of this task we address the problem of distinguishing between multiple speakers in a noisy and reverberant environment using a pair of microphones. Our approach is based on tracking azimuth angle measurements for every person. When the speakers' paths cross, the azimuth features will not have enough resolution to distinguish between the people. In order to disambiguate such difficult cases, we attempt to combine azimuth cues with speaker-specific voice features. We also assume that only one of the speakers is active at a time.

We consider a probabilistic framework, where every speaker is described with a latent state variable that includes the “true” azimuth angle and formant frequencies. Given a segment of the input signals, we update our beliefs about states of all speakers. The beliefs become a basis associating the input segments with one of the speakers. Our implementation applies a sample based version of the



Fig. 1. Philips iCat interface robot [6]. The two microphones are mounted on the sides of the bottom panel.

joint probabilistic data association filter to compute the interesting association probabilities.

For a robot, it is difficult to match the hearing capabilities of a human. Human auditory system finds spatial speaker locations by taking into account various physical effects, like the acoustic shadowing created by the head and sound reflections caused by the outer ear ridges. This ability, together with the sophisticated mechanism for voice recognition allows for accurate speaker identification by humans.

Several methods have been developed in order to mirror the human sound source location capability. Typically, one relies on time-difference of arrival (TDOA) between signals from a pair of microphones. The TDOA measurements indicate the azimuth angle between the source location and the geometrical center of the microphone pair [18]. The sensing resolution can be increased by using multiple microphone pairs [14].

Speaker identity is correlated with the physiological and behavioral characteristics of the speaker. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics and dynamic features spanning several segments). The most common are the features related to the

spectral envelope. In the simplest case these features are taken to be the LPC (linear predictive coding) coefficients. More elaborate approaches use LPC-cepstral coefficients together with their regressions coefficients [13]. A comprehensive survey of speaker identification techniques is provided in [11].

II. SENSOR FEATURES

As already mentioned, our robot is equipped with a pair of horizontally aligned microphones (Fig. 1). We will denote the left and right microphone signals as $m^\ell(t)$ and $m^r(t)$, where t is a discrete time step. Our algorithm operates on fixed-length segments representing 25ms-long windows from the input signals. We will use $n = 1, 2, \dots$ as a segment index.

From every segment we will compute a set of features $\lambda_n = [\theta_n, \gamma_n, \phi_n]$, where θ_n are azimuth angle measurements, γ_n is a discrete vowel indicator, and ϕ_n is a vector of vowel formant frequencies. In the rest of this section, we present the details of computing these features.

A. Azimuth features

Our robot will operate in office-like places with moderate acoustic noise levels, and where the surfaces reflect the source signal. Therefore we will consider the signals in the frequency domain, since it allows for easier removal of noise and reverberation artifacts.

To extract the relative delay between the input signals, we use the Generalized Cross-Correlation (GCC). GCC is just the inverse Fourier Transform of the received signal cross-power spectrum scaled by a weighting function. The cross-correlation between the left and right signal is

$$R_n(\tau) = \int_{-\infty}^{\infty} G_n(\omega) F_n^\ell(\omega) F_n^{r*}(\omega) \exp(j\omega\tau) d\omega, \quad (1)$$

where F_n^ℓ and F_n^r denote the Fourier transforms of the n th input segments, and $*$ denotes conjugate. We use the phase transform (PHAT) [5] as the pre-filtering weight $G_n(\omega)$:

$$G_n(\omega) = |F_n^\ell(\omega) F_n^{r*}(\omega)|^{-1}.$$

Such a weighting function places equal importance on each frequency by dividing the spectrum by its magnitude. As a result, the filter preserves only information about the phase differences between the signals. We find the five delays τ that yield the highest correlation in (1). These delays are then transformed into azimuth angles, and stored as a measurement vector $\theta_n = [\theta_{n,1}, \dots, \theta_{n,5}]$ for the n th segment. In this way we deal with possible false-positives – spurious maxima in the correlation function.

B. Voice features

For simplicity, we have chosen the LPC-based features to represent speaker-specific vocal tract characteristics. For every segment of the input signal we estimate the spectral envelope (using standard LPC [4]) and decide whether the

envelope represents a vowel. For the detected vowel we find the three highest peaks in the envelope and consider those as the formant frequencies.

Each segment is first analyzed on basis of its energy distribution. If the energy is mainly concentrated in the lower regions of the spectrum (0–4kHz) then the formant frequencies of the segment are estimated from peaks in the LPC spectrum. In order to identify a possible vowel present in the segment, the found formant frequencies are compared with an table of mean formant frequency locations and a table of average formant bandwidths [4]. Since occasionally the peaks may occur due to noise, a positive identification is verified with a sub-band filter. Essentially, the filter aims to track the detected formants to verify their consistency over the segment.

In the rest of the paper, we use $\gamma_n \in \{/i/, /I/, /E/, /@/, /a/, /c/, /U/, /u/, /A/, /R/\}$ to denote the detected vowel from segment n . Where there is no vowel detected we will write $\gamma_n = 0$. When $\gamma_n \neq 0$, we will denote the 3 measured formant frequencies as a vector $\phi_n = [\phi_{n,1}, \phi_{n,2}, \phi_{n,3}]$.

III. MODEL

A. Overview

Our primary goal is association of the measured features λ_n with one of the speakers. For this purpose, we describe the i th speaker with a state variable \mathbf{s}_n^i , which summarizes the persons' location and voice properties during the n th segment. Since the state cannot be directly observed, we will consider it as a hidden (latent) random variable with a prior distribution $p(\mathbf{s}_0^i)$. We will also define a *sensor model*, which describes a probabilistic dependency of measurements on the state $p(\lambda_n | \mathbf{s}_n^i)$. Under such a framework will compute posterior state distribution $p(\mathbf{s}_n^i | \lambda_{1:n})$ given measurements using Bayesian filtering [8]. On the basis of this distribution we associate segments with speakers.

The state of i th person during the n th segment is described by $\mathbf{s}_n^i = [\mathbf{z}_n^i, \mathbf{f}^i]$, where $\mathbf{z}_n^i = [y_n^i, \dot{y}_n^i, x_n^i, \dot{x}_n^i]$ is a 4-dimensional vector denoting the position and speed in the usual Cartesian coordinates, and \mathbf{f}^i is a “formant profile” of the person. The profile is a collection of 15 characteristic formant frequencies of the detected vowels $\mathbf{f}^i = [f_{/i/}^i, \dots, f_{/R/}^i]$. Each f_γ^i represents the three formants for the vowel γ . We assume that the profiles are constant, therefore we did not use subscript n with \mathbf{f}^i .

We set the center (0, 0) of the coordinate system in the middle of the microphone pair. Note, that we cannot measure the distance between the speaker and microphones. Thus, in the (x, y) coordinates, we will be effectively estimating the ratio x/y from the azimuth data. Our choice for (x, y) coordinates, follows from the fact that we can now apply a well-behaved Langevin motion model.

B. Prior

The prior state distribution summarizes our knowledge about states before the measurements become available. We

will factorize this distribution into a product of Gaussian (Normal) density functions

$$p(\mathbf{s}_0^i) = \mathcal{N}(\mathbf{z}^i | \mathbf{m}_z, \mathbf{R}_z) \mathcal{N}(\mathbf{f}^i | \mathbf{m}_f, \mathbf{R}_f) \quad (2)$$

We assume that a-priori every person is standing still at the front of the robot, $\mathbf{m}_z = [1, 0, 0, 0]$. In the experiments we have chosen a diagonal covariance $\mathbf{R}_z = \mathbf{I}_{4 \times 4}$. The mean vector \mathbf{m}_f is chosen using a table of typical (averaged of male and female speakers) formant frequencies for the interesting vowels [4].

C. Langevin motion model

The location of person may change continuously. For simplicity, we assume a quasi-static location within each segment. Our segments correspond to 25ms intervals, and we do not expect the speakers to move substantially within such intervals. The motion between the segments is described as a stochastic Langevin process [18]

$$x_n = x_{n-1} + \delta \dot{x}_n \quad (3)$$

$$\dot{x}_n = \alpha \dot{x}_{n-1} + \beta \nu, \quad (4)$$

where $\nu \sim \mathcal{N}(0, \sigma_\nu)$ is a stochastic velocity disturbance, α and β are coefficients (see experiments) and δ denotes the time gap between segments. Identical equations hold for the y coordinate.

D. Sensor model

Sensor model defines a probabilistic dependency between the state of a person \mathbf{s}_n^i and the measured quantities $\lambda_n = (\theta_n, \gamma_n, \phi_n)$. This model is the same for every person, so we omit the superscript i . The model takes the form

$$p(\lambda_n | \mathbf{s}_n) = p(\gamma_n) p(\phi_n | \gamma_n, \mathbf{f}) p(\theta_n | \mathbf{z}_n), \quad (5)$$

$$p(\theta_n | \mathbf{z}_n) = \frac{1/5}{\sqrt{2\pi}\sigma} \sum_{k=1}^5 \exp \frac{(\tan(\theta_{n,k}) - x_n/y_n)^2}{\sigma^2} \quad (6)$$

$$p(\phi_n | \gamma_n, \mathbf{f}) = \mathcal{N}(\phi_n | \mathbf{f}(\gamma_n), \mathbf{R}_s(\gamma_n)) \quad \text{iff } \gamma_n \neq 0, \quad (7)$$

where the $p(\gamma_n)$ is chosen uniform. For the position measurements, we use a mixture of Gaussians, each centered at one of the measured hypothetical azimuth angles. The constant $1/5$ ensures a proper normalization of the mixture. For simplicity, we assume Gaussian density for the measured formants given the ‘‘true’’ formant profile. The term $\mathbf{f}(\gamma)$ denotes entries from \mathbf{f} corresponding to vowel γ , and \mathbf{R}_s is a (diagonal) sensor noise variance. When there was no vowel detected, i.e. $\gamma_n = 0$, we use a uniform density in place of $p(\phi_n | \gamma_n, \mathbf{f})$.

IV. FILTERING

In this section we describe a procedure that updates state distributions $p(\mathbf{s}_n^i | \lambda_{1:n})$ from a sequence measurements $\lambda_{1:n}$. These distributions represent our knowledge about the motion and formant profile of each speaker. Given a new measurement, we can compute association probabilities

to find the speaker that is the most likely source of the measurement.

We have formulated our problem as a stochastic time-series process with noisy observations [8]. The interesting distributions can be computed with a recursive Bayesian filtering procedure. However, our task is an instance of a more general class of probabilistic multi-target tracking problems. Exact filtering for these problems is typically intractable since one has to couple state estimation with measurement-target association. Within the Bayesian framework, a well-established approximate approach for dealing with association uncertainty is by joint probabilistic data association filters (JPDAFs) [3].

Here, we apply the JPDAF scheme together with sample-based representation of the motion component. This component will be estimated using the Sampling-Importance-Resampling (SIR) [7]. The importance weights will be modified in order to account for measurement-target association uncertainty, as proposed in [16]. On the other hand, the formant component will be represented with a Gaussian family. Since both the prior and the sensor models are Gaussian, this component can be seen as a linear Gaussian system with data association uncertainty [8].

A. Representation

For simplicity, the filtered distribution on the state of the i th speaker after the n th segment will be approximated as $p(\mathbf{s}_n^i | \lambda_{1:n}) \approx p(\mathbf{z}_n^i | \lambda_{1:n}) p(\mathbf{f}^i | \lambda_{1:n})$. The factorial formula is an approximation, since in general the motion component \mathbf{z}_n^i and formant profile \mathbf{f}^i will not be independent given the measurements.

Given the highly non-linear and multi-modal sensor model for location measurements (6), we choose a particle-based representation of the motion component

$$p(\mathbf{z}_n^i | \lambda_{1:n}) \approx \sum_{k=1}^M \delta(\mathbf{z}_n^i - \mathbf{z}_{n,k}^i),$$

where M is the number of particles (per object), $\mathbf{z}_{n,k}^i$ is the k th particle, and $\delta(\cdot)$ is a delta function.

The distribution of the formant profiles will be approximately represented with a Gaussian density function

$$p(\mathbf{f}^i | \lambda_{1:n}) \approx \mathcal{N}(\mathbf{f}^i | \mathbf{m}_n^i, \mathbf{R}_n^i),$$

where \mathbf{m}_n^i and \mathbf{R}_n^i are the mean and covariance. The prior (2) and sensor (7) models assume diagonal covariances, therefore \mathbf{R}_n^i will also be diagonal.

B. Algorithm

Figure 2 presents the overview of the filtering algorithm for our problem. For every segment, it comprises three basic steps: 1) prediction of states from past data, 2) computing association probabilities, 3) updating the states with the current measurement. Below we describe these steps in detail.

For each segment n : Compute measurement λ_n Compute predictive densities $p(\mathbf{s}_n^i \lambda_{1:n-1})$ for all i Compute association probabilities $p(\beta_n = i)$ for all i Compute updated densities $p(\mathbf{s}_n^i \lambda_{1:n})$ for all i
--

Fig. 2. Overview of the filtering algorithm.

1) *Prediction*: Assume that at $n-1$ there are I speakers, and their state distributions are parametrized by $\mathbf{z}_{n-1,k}^i$, \mathbf{m}_{n-1}^i and \mathbf{R}_{n-1}^i . Predictive distribution for the motion component, follows from the standard SIR scheme [7], where we obtain predictive particles $\hat{\mathbf{z}}_{n,k}^i$ by sampling from the motion model conditioned on $\mathbf{z}_{n-1,k}^i$. The formant profiles are assumed constant, so we just use the current distribution

$$p(\mathbf{s}_n^i | \lambda_{1:n-1}) = \sum_{k=1}^M \delta(\mathbf{z}_n^i - \hat{\mathbf{z}}_{n,k}^i) \mathcal{N}(\mathbf{f}^i | \mathbf{m}_{n-1}^i, \mathbf{R}_{n-1}^i),$$

We also predict the state of a new $(I+1)$ th speaker, by setting the predictive distribution equal to the prior (2).

2) *Association events*: Let $\beta_n = i$, denote the event that the i th speaker was the source of the measurement $\lambda_n = (\theta_n, \gamma_n, \phi_n)$. The event $\beta_n = I+1$ corresponds to a new speaker. We can compute association probabilities as

$$\beta_n^i = p(\beta_n = i) = \alpha \eta_i \sum_{k=1}^M p(\theta_n | \hat{\mathbf{z}}_{n,k}^i),$$

$$\eta_i = \mathcal{N}(\phi_n | \mathbf{m}_{n-1}^i(\gamma_n), R + \mathbf{R}_{n-1}^i(\gamma_n))$$

where α is a constant, ensuring that $\sum_{i=1}^{I+1} \beta_n^i = 1$. The term η_i indicates how the measured formants ϕ_n of the vowel γ_n fit to the i th profile, where $\mathbf{m}(\gamma)$ indicates a subset of entries from \mathbf{m} that correspond to the vowel γ . (Similarly for \mathbf{R}_{n-1}^i). If there was no vowel detected $\gamma_n = 0$ we set $\eta_i = 1$.

The association probabilities allow to find the most likely speaker by taking $\text{argmax}_i p(\beta_n = i)$. We can also decide whether there is a new speaker in the environment by evaluating $p(\beta_n = I+1)$. In our implementation we decided to introduce a new speaker only if $p(\beta_n = I+1) > 0.9$.

3) *Update*: The SIR filter updates the sampled-based distribution of the motion component by weighting the predictive particles. The weight of particle $\hat{\mathbf{z}}_{n,k}^i$ is

$$w_{n,k}^i = \alpha \beta_n^i p(\theta_n | \hat{\mathbf{z}}_{n,k}^i),$$

where α is a constant ensuring that $\sum_{k=1}^M w_{n,k}^i = 1$. Note the term β_n^i which accounts for association uncertainty. After computing the weights, we obtain a new set of particles $\mathbf{z}_{n,k}^i$ with the standard importance resampling.

If there was no vowel detected in the formant predictive densities do not require the update step, and are propagated unchanged. Otherwise, we update only the profile of the most likely speaker $j = \text{argmax}_i p(\beta_n = i)$. Once the association has been resolved, the update of the Gaussian

formant density is identical to the update step in standard linear Gaussian models [8]. We note, that updating only the most likely source is a simplification since it does not take the association ambiguity into account. In addition we found that $\eta_i = \mathcal{N}(\phi_n | \mathbf{m}_{n-1}^i(\gamma_n), R)$ yielded more robust association assignments. We believe that this is due to the differences in the number of received vowel examples for each object. These differences result in different covariances for the found formant frequencies for each speaker. Since the vowel profiles are learned online, in a very limited amount of time, the differences in covariance can become considerable. By using the same covariance R for every speaker, we guarantee that the speakers who talk frequently will not be favored over the speakers who talk infrequently.

V. EXPERIMENTS

We demonstrate our approach using a collection of stereo recordings obtained with the Philips iCat interface robot. We first show the single-speaker tracking scenario. Next, we consider various two-speaker configurations, where the recorded signals contain short sentences (approx. 1s–2s) produced by the speakers in turns. Both speakers were male. In total, each of our signals was approx. 45s long.

The recordings were taken in a rectangle-shaped office room with moderate noise conditions (PC running in the background). Due to the strongly reflective walls and floor of the room, we expect that reverberation will be the main cause of the clutter. The microphones were placed near the longer wall (approx. in the middle), 1.5m above the ground level. The distance between the microphones was 0.3m.

A. Parameters

We set the model parameters as follows: The sensor variance for formant measurements \mathbf{R}_s is a diagonal 15×15 identity matrix, $\mathbf{R}_s = \mathbf{I}_{15 \times 15}$. The sensor variance for azimuth angle measurements was chosen as $\sigma^2 = 0.1$. The parameters for the motion model (the Langevin) process were $\alpha = 0.8$, and $\beta = 0.6$. These values correspond to a human (slowly) walking in a room. For the SIR filter we used $M = 5000$ (in the two speaker experiments we used $M = 2000$) samples per object.

B. Single speaker

Figure 3 presents the measured azimuth angles (three for every segment) and the estimated trajectory of a single speaker. The thin line gives the estimated expected location of the speaker (expressed as an azimuth angle). In fact the speaker started in front of the robot and moved to the side for about half the signal duration. Then the speaker stopped about 3 meters away from the microphones. Therefore in the second part of the signal, the angle measurements are much less reliable.

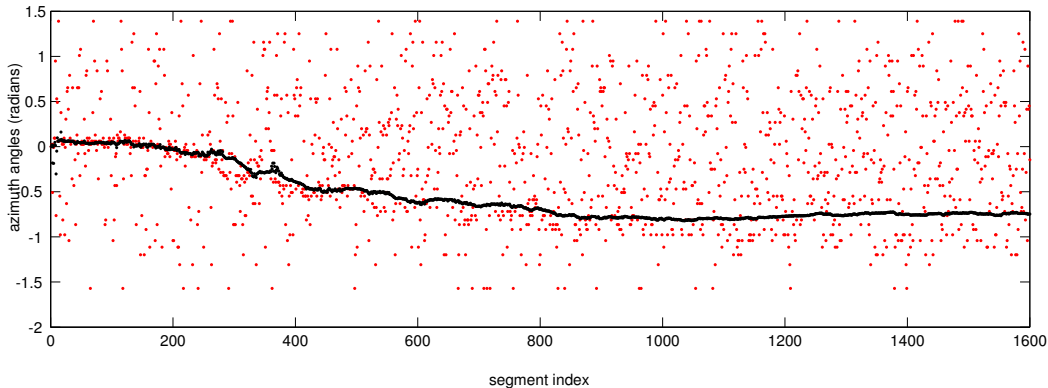


Fig. 3. Tracking of a single person. The measured azimuth angles are overlaid with the estimated expected location of the speaker.

C. Two speakers

In this experiments we consider two speakers who are initially positioned on the opposite sides of the robot. In order to build profiles that can be compared, the speakers were requested to speak short sentences that contained the vowels mentioned in section II-B. The speakers gradually move toward each other, while decreasing their distance from the microphones. In Fig. 4 we present the azimuth measurements (as thin dots) and the expected locations of the target speakers (in bold line). When both speakers are closely in front of the microphones the tracker is able to successfully discriminate between the speakers. Despite the strong clutter, the tracker can estimate the location quite reliable for the most part of the sequence. In the final part, when the targets move further away from the microphones, the tracker loses one of these targets.

With the use of only two microphones we cannot determine if the speakers are in front of the microphone pair or not. Furthermore, with the use of only two microphones we receive imprecise readings whenever the speakers are in the same axis as the microphone pair. Therefore the speakers were requested to remain in front of the microphone pair.

D. TDOA estimation

Finally, we show how estimation of the vowel formants improves the TDOA measurements. Typically, before further processing the PHAT function applies a low-pass filter (8kHz cut-off). This filter removes high-frequency noise present in the signal, and preserves the speech components, which are located in the lower-end of the spectrum. When the vowel formants are available, we can now use the knowledge about vowel spectral location to more precisely select the filter cut-off. In Fig. 5 we show the cross-correlation vs. delay plot. In the left panel we have used a fixed low-pass filter. In the right panel, the same plot obtained with a low-pass filter, where the cut-off was selected to be the third formant frequency. In this way we could extend the range of discarded frequencies and remove many spurious peaks in the correlation function.

VI. CONCLUSIONS

We have presented a system that allows a (static) robot to keep track of multiple speakers in its neighborhood. This work is a part of a larger effort that aims for providing a voice-based robot interface. Our approach relies on two microphones, which provide azimuth angle cues about locations of the speakers around the robot. The azimuth measurements are combined with a “formant profile” of each person. The profiles represent intrinsic speaker properties, which are learned on-line. Although the presented test involved a limited number of tracked speakers, they already indicate the benefits of using vowel-intrinsic features.

The presented ideas can be extended to more elaborate scenarios, where the audio cues are used jointly with visual sensors. As an example, robot could use the audio signals to steer the camera toward the current speaker. Alternatively, in limited closed areas, the audio feature could help keep track of a person who disappears from robots field of view.

ACKNOWLEDGMENTS

We thank Albert van Breemen of Philips Research for providing the iCat microphone system for our experiments.

The work of Ben Kröse described in this paper was conducted within the EU Integrated Project COGNIRON (“The Cognitive Companion”) and was funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020. Wojciech Zajdel was supported by the Technology Foundation STW (grant no ANN.5312), applied science division of NWO and the Dutch Ministry of Economic Affairs.

REFERENCES

- [1] Arras, K.O., Philippsen, R., Tomatis, N., de Battista, M., Schilt, M. and Siegart, R. “A Navigation Framework for Multiple Mobile Robots and its Application at the Expo.02 Exhibition”, in Proceedings of the IEEE International Conference on Robotics and Automation (2003).
- [2] H. Asoh, N. Vlassis, Y. Motomura, F. Asano, I. Hara, S. Hayamizu, K. Itou, T. Kurita, T. Matsui, R. Bunschoten, and Ben Kröse. Jijo-2: An office robot that communicates and learns. *IEEE Intelligent Systems*, 16(5):46–55, Sep/Oct 2001.

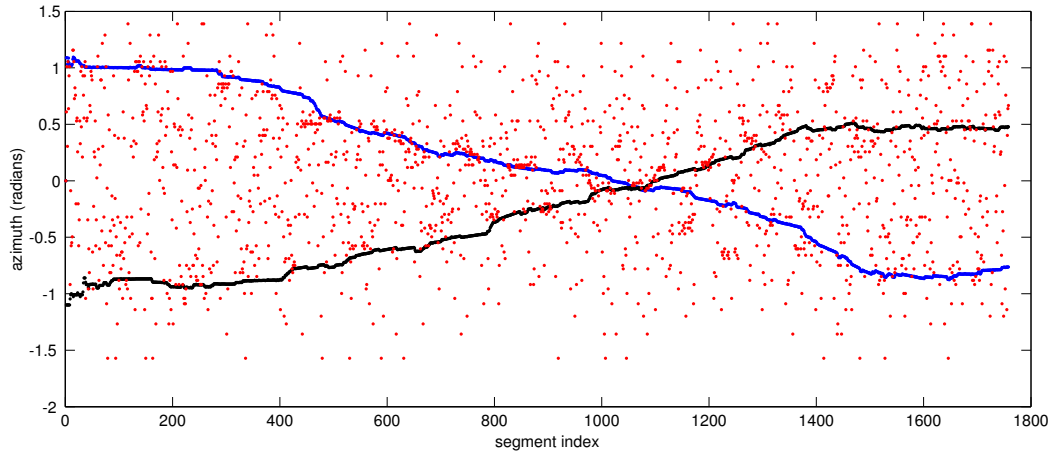


Fig. 4. Tracking two speakers. The horizontal axis represents segments from the input signals. The vertical axis shows location as azimuth angles (in radians). Each dot represents a measured angle. Two bold lines give the estimated speakers' locations.

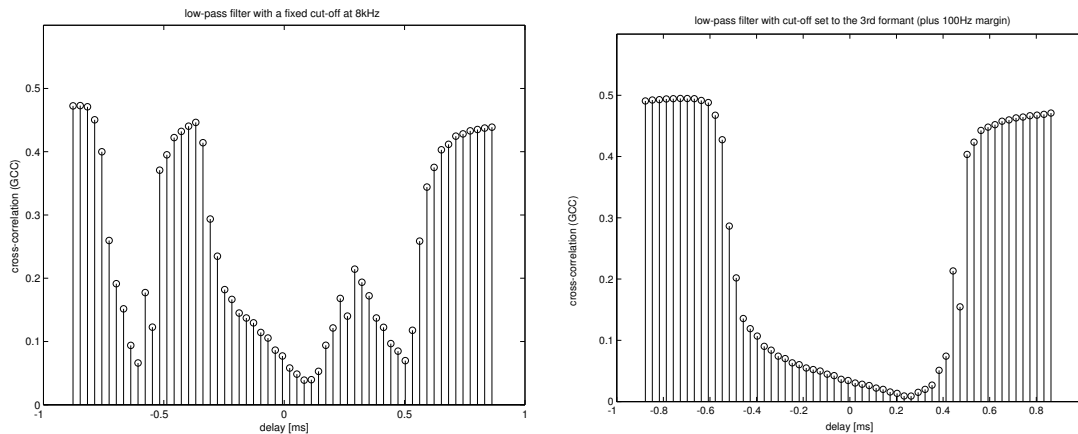


Fig. 5. Comparison of TDOA measurements obtained from regular PHAT algorithm against measurements obtained using vowel-specific frequency range.

- [3] Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, 1988.
- [4] J.R. Deller, Jr., J.H.L. Hansen and J.G. Proakis. *Discrete Time Processing of Speech Signals*. IEEE Press, 2000.
- [5] M. Brandstein, H. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1996.
- [6] A.J.N. van Breemen. Animation Engine for Believable Interactive User-Interface Robots. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2004), 2004.
- [7] A. Doucet, N. de Freitas, and N. Gordon (eds.), *Sequential monte carlo methods in practice*, Springer-Verlag, 2001.
- [8] Z. Ghahramani. Learning dynamic Bayesian networks. In C.L. Giles and M. Gori, editors, *Adaptive Processing of Temporal Information*, Lecture Notes in Artificial Intelligence, pages 168–197. Springer-Verlag, 1998.
- [9] D. Schulz, W. Burgard, D. Fox, and A.B. Cremers. People Tracking with a Mobile Robot Using Sample-based Joint Probabilistic Data Association Filters. *Int. Journal of Robotics Research*, 22 (2), 2003.
- [10] J. Fritsch, M. Kleinhagenbrock, S. Lang, T. Btz, G. A. Fink, and G. Sagerer, Multi-modal anchoring for human-robot-interaction. *Robotics and Autonomous Systems, Special issue on Anchoring Symbols to Sensor Data in Single and Multiple Robot Systems*, 43(2–3):133–147, 2003.
- [11] S. Furui. Research on individuality features in speech waves and automatic speaker recognition techniques. *Speech Communication*, 5(2):183–197, 1986.
- [12] A. J. Davison, M. Montemerlo, J. Pineau, N. Roy, S. Thrun and V. Verma, Experiences with a Mobile Robotic Guide for the Elderly, in Proc. of the AAAI National Conf. on Artificial Intelligence (2002)
- [13] T. Matsui and S. Furui. A text-independent speaker recognition method robust against utterance variations. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1992.
- [14] M. Omologo, P. Svaizer. Talker Localization and Speech Enhancement in a Noisy Environment using a Microphone Array based Acquisition System. EUROASPEECH 1993.
- [15] A. Swain, W.H. Abdulla. Estimation of LPC Parameters of Speech Signals in Noisy Environment. In *IEEE TENCON*, 2004.
- [16] D. Schulz, W. Burgard, D. Fox, and A. B. Cremers. Tracking multiple moving targets with a mobile robot using particle filters and statistical data association. In *IEEE Int. Conf. on Robotics and Automation*, 2001.
- [17] S. Thrun, M. Bennewitz, W. Burgard, A.B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C.R. Rosenberg, N. Roy, J. Schulte and D. Schulz, MINERVA: A Tour-Guide Robot that Learns, {KI} - Kunstliche Intelligenz, (1999) 14-26
- [18] J. Vermaak and A. Blake. Nonlinear filtering for speaker tracking in noisy and reverberant environments. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2001.