Auditory Inspired Binaural Robust Sound Source Localization in Echoic and Noisy Environments

Martin Heckmann, Tobias Rodemann, Frank Joublin, Christian Goerick, Björn Schölling

2006

Auditory Inspired Binaural Robust Sound Source Localization in Echoic and Noisy Environments

Martin Heckmann, Tobias Rodemann, Frank Joublin, Christian Goerick
Honda Research Institute Europe GmbH
Carl-Legien-Strasse 30
D-63073 Offenbach/Main, Germany
{*martin.heckmann, tobias.rodemann, frank.joublin, christian.goerick*}*@honda-ri.de*

Björn Schölling
Institut für Automatisierungstechnik
Technische Universität Darmstadt
D-64283 Darmstadt, Germany
bjoern.schoelling@rtr.tu-darmstadt.de

Abstract— **We propose a new approach for binaural sound source localization in real world environments implementing a new model of the precedence effect. This enables the robust measurement of the localization cue values (ITD, IID and IED) in echoic environments. The system is inspired by the auditory system of mammals. It uses a Gammatone filter bank for preprocessing and extracts the ITD and IED cues via zero crossings (IID calculation is straight forward). The mapping between the cue values and the different angles is learned offline which facilitates the adaptation to different head geometries. The performance of the system is demonstrated by localization results for two simultaneous speakers and the mixture of a speaker, music, and fan noise in a normal meeting room. A real time demonstrator of the system is presented in [1].**

I. Introduction

Sound source localization for a robot is an important yet difficult task. In real environments the noise generated by the robot itself as well as the echoes disturb the localization process. Furthermore, in addition to the direct path also the reflections from the walls and the furniture impinge on the microphones. The reflections arrive from different directions than the actual signal and therefore interfere with the localization of the source.

Most systems for source localization are based on an autocorrelation. In order to deal with echos they perform a weighting of the correlation function [2] or select measures based on a reliability criterion [3], [4]. A different approach to overcome the echos is inspired by psychoacoustics, more precisely the *Precedence Effect*, and only uses the onsets of the signals to measure the localization cues[3].

Since the task gets easier as the number of microphones and their distance is increased a multitude of systems uses arrays of microphones[5], [6]. For sound source localization on a robot like Asimo the dimensions of the robot restrict the size of the array and therefore make the problem more difficult. Furthermore biological systems are still far better in localizing sound sources in noisy environments than technical systems and therefore better performance for technical systems which try to understand and implement solutions found in biology can be expected. For these reasons we are investigating binaural source localization. The number of systems performing binaural localization is much more limited [7], [8], [9] especially of those which work in echoic environments [10].

Binaural systems commonly work in the frequency domain (either via FFT or as in our case by using a band pass filterbank) and use the following cues:

Interaural Time Difference (ITD): The time delay between the left and right signal.

Interaural Intensity/Level Difference (IID/ILD): The intensity difference between the left and right signal.

Interaural Envelope Difference (IED): The time delay between the left and right envelope modulations.

These cues are known to be also responsible for the sound source localization capabilities of humans [11].

In the following we will first detail our echo suppression mechanism based on the Precedence Effect which enables robust measurements in echoic environments. Then we introduce our basic localization system and finally present some results.

II. Modeling the precedence effect

It is known that the Precedence Effect makes localization in echoic environments possible for humans. The main findings are [11]:

1) A leading sound suppresses localization of a shortly following sound ($\approx 40\,\mathrm{ms}$).
2) The lagging sound still has a small influence on the localization of the leading sound.
3) A lagging sound sufficiently more intense than the leading sound ($10 - 15\,\mathrm{dB}$) overwrites the precedence effect.
4) The precedence effect takes some time to build up and is influenced by a change in the acoustical environment.
5) Despite the precedence effect information on the echoes (room size) is still available to the listeners.

For modeling the precedence effect 1) and 3) are of special interest. The most basic model is to perform the localization only in the onsets of a signal and inhibit following onsets for a fixed time span determined a priori [3]. The motivation

1-4244-0259-X/06/$20.00 ©2006 IEEE

$$x_s(k) = \begin{cases} 0 & k = 0 \\ x(k) & x_s(k-1) \leq x(k) \wedge k > 0 \\ (1 - 1/\tau) \cdot x_s(k-1) + 1/\tau \cdot x(k) & x_s(k-1) > x(k) \wedge k > 0 \end{cases} \quad (1)$$

$$x_s(k) = \begin{cases} 0 & k = 0 \\ x(k) & x_s(k-1) \leq x(k) \wedge k > 0 \\ x(k) \cdot \vartheta & x_s(k-1) > x(k) \wedge x_s(k-1) \leq x(k-1) \wedge k > 0 \\ (1 - 1/\tau) \cdot x_s(k-1) + 1/\tau \cdot x(k) & x_s(k-1) > x(k) \wedge x_s(k-1) > x(k-1) \wedge k > 0 \end{cases} \quad (2)$$

behind this is that with the onsets only the direct path is captured and the measurement is stopped when the echoes arrive and thus implements 1). In our model we also included 3) such that a loud signal triggers again the measurement process even if the inhibition time is not over. Additionally we changed the measurement point and do not use the onsets of the signal but the maxima. A first reason for doing so is that the onsets are difficult to determine reliably and a threshold is necessary to make the decision if the current rise in energy is really an onset or just noise. Secondly we made the observation that the cues used for localization are rather unstable at the onsets, stabilize until the maximum and then are affected by the echoes in the part after the maximum. The cues in smaller maxima following a maximum at the signal onset are dominated by the echoes. Therefore we implemented an inhibition of shortly following smaller maxima. For doing so a nonlinear smoothing of the signal envelope was developed. It acts in two modes. In the first mode the smooth envelope $x_s(k)$ rises with the signal envelope $x(k)$. When the signal envelope changes from the rising phase to a falling phase, hence after a maximum, the smoothing changes its mode and now performs a smoothing of the envelope signal with a first order Infinite Impulse Response (IIR) filter. When the smooth signal falls below the envelope signal the smoothing changes again in its rising phase. As a consequence the onsets are conserved and the signal is only smoothed after the onsets. A measurement point for the localization cues is generated one sample before the change from the rising to the falling phase and hence at the maxima of the signal (compare Eq. 1 where $x(k)$ is the original envelope signal, $x_s(k)$ the resulting smooth envelope, and τ the time constant of the IIR filter). For the calculation of the envelope we use a rectification and low-pass filtering. The result of this smoothing can be seen in Fig. 1. Due to the smoothing after the maxima the maximum at $0.17\,\mathrm{s}$ does not produce a measurement point as at this point the smooth signal is still higher than the envelope signal and hence the smoothing does not change its mode. In contrast the maximum at $0.09\,\mathrm{s}$ is much higher than the one at $0.03\,\mathrm{s}$ and therefore the smoothing changed its mode already well before and a measurement is generated at $0.09\,\mathrm{s}$. In order to also inhibit only slightly stronger maxima following shortly after a maximum we introduced an additional inhibition factor ϑ in the smoothing process. At the measurement point the smooth signal is multiplied with this inhibition factor and therefore raised to a higher value from which it then falls again in the following smoothing phase (compare Eq. 2). The impact of the additional inhibition is shown in Fig. 2. As can be seen the smooth signal now jumps at the measurement points at $0.03\,\mathrm{s}$ and $0.09\,\mathrm{s}$. Due to this jump the smooth signal is at $0.13\,\mathrm{s}$ still above the envelope signal and hence the smoothing mode does not change and no measurement point is generated. Thus this inhibition factor leads to the inhibition of the measurement at $0.13\,\mathrm{s}$ where the localization cues are normally considerably affected by echoes. A single sound event procudes a maximum in the left and right channel. Therefore if maxima in the two channels are closer together than $40\,\mathrm{ms}$ only the earlier maximum is kept. All in all our model of the precedence effect leads to measurements only in the initial rising part of the signal, hence the first wavefront. Measurements in shortly following maxima are inhibited except if their amplitude is significantly higher than the leading maximum. The time the inhibition is active depends on the shape of the signal. The faster the signal falls the shorter the inhibition time. This is another marked difference to previous onset based localization systems which

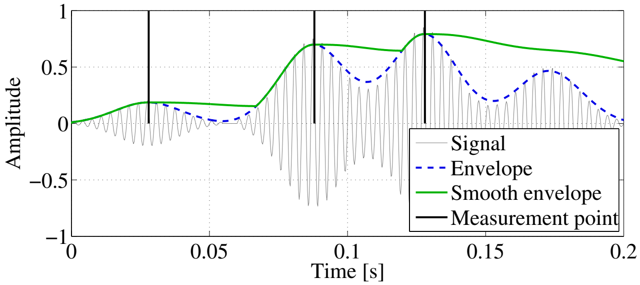

Fig. 1. The procedure of the nonlinear smoothing is visualized. The smooth envelope rises with the original envelope and falls with a time constant. As a consequence the maximum at $0.17\,\mathrm{s}$ is inhibited whereas the maximum at $0.09\,\mathrm{s}$ produces a measurement.

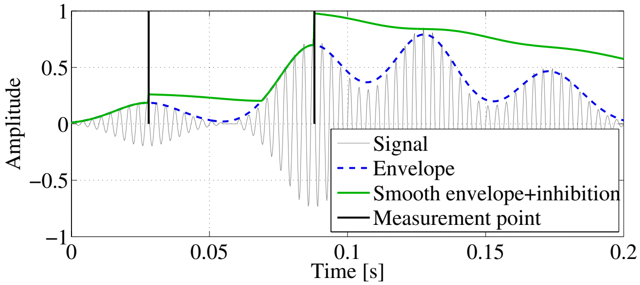

Fig. 2. The additional inhibition factor makes the smooth signal jump at the measurement points which leads also to the inhibition of the maximum at $0.13\,\mathrm{s}$.

perform an inhibition for a fixed time.

III. Basic System Architecture

Instead of the real Asimo head we used a dummy head for the results presented here. Microphones were attached to the ears of the head. In line with our biology inspired approach we first apply a Gammatone filter bank [12] to the input signals. Stationary noise was estimated in the beginning of the signals and then removed via spectral subtraction. In Figure 3 an overview of the complete system is given.

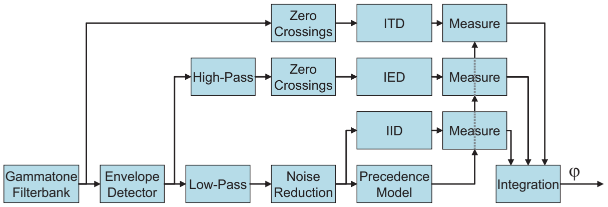

Fig. 3. System Overview

A. Cue Extraction

We use zero crossings to extract the ITD instead of the autocorrelation. Zero crossings are robust when applied to bandpass signals, significantly faster to calculate than an autocorrelation and biologically more plausible [13], [14], [15]. The ITD is measured at each zero crossing and then kept at this value until the next zero crossing occurs. For the IID values we calculated

$$IID(c,k) = \frac{x_L(c,k) - x_R(c,k)}{\max\left(x_L(c,k), x_R(c,k), x_{Min}\right)} \ , \qquad (3)$$

where $x_L(c,k)$ and $x_R(c,k)$ are the envelope signals of the left and right channel after noise reduction at sample k and frequency channel c and x_{Min} the minimal expected signal level which prevents divisions by zero. To obtain the IED values the envelope signal was high pass filtered. The resulting signal contains only the modulations of the envelope resulting from *unresolved harmonics* [11]. The time difference between these modulations in the left and right channel is again measured via zero crossings. In our implementation the IED cue does not produce reliable measurements and we therefore will only detail the ITD and IID cues in the remaining of the paper.

The cues are evaluated at the time defined by the non-linear envelope smoothing. Based on the found maxima a $10\,\mathrm{ms}$ long measurement window is formed. In the current implementation the measurement window starts $13\,\mathrm{ms}$ before the maximum and ends $3\,\mathrm{ms}$ before the maximum. The final cue value for this channel and instance in time is the mean of the cue value in the window.

B. Mapping Matrix Calculation

As the geometry of the artificial head used is rather complex there is no straight forward mapping between the cue values and the corresponding angles possible. We therefore learn this mapping in an offline procedure. Sounds from known directions are presented to the head and localization cue values are extracted. An average cue value for a given location and a given frequency channel can be calculated from the calibration data. With this average value a mapping between the cue value and the angle can be established. The result of this mapping is visualized in Fig. 4. We used 25 azimuth positions ranging from $-90°$ to $90°$ with $10°$ increment and a reduced increment around $0°$ in order to increase resolution around $0°$. The localization is limited to $-90°$ to $90°$ azimuth as we currently do not use combinations of cues or spectral characteristics of the signals to perform a front/back decision or elevation estimation. The ITD and IID cues per se can not resolve this when using only two microphones.

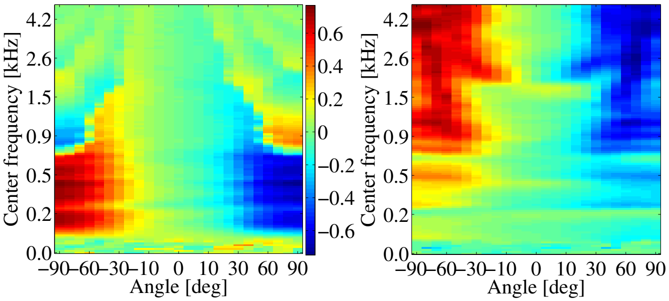

Fig. 4. The mapping matrices learned are visualized. On the x-axis the azimuth angle and on the y-axis the center frequency of the channel under consideration is given. As can be seen ITD (left plot) works well for low frequencies and IID (right plot) better for high frequencies. For frequencies above $0.8\,\mathrm{kHz}$ ITD gets ambiguous which means that there are identical ITD values for different angles.

C. Frequency Dependent Cue Confidence

From the data used in the mapping matrix calculation the variances $\sigma_{\mathrm{ITD}}(c,\varphi)^2$ for the three cues at a given channel c and angle φ can be calculated. Based on these variances and the average cue values $\mathrm{ITD}_M(c,\varphi)$ a confidence value for each cue at each channel averaged over all directions M is calculated[1]:

$$\eta_{\mathrm{ITD}}(c) = \sqrt{\frac{1}{M}\sum_{\varphi=-90°}^{90°}\frac{\mathrm{ITD}_M(c,\varphi)^2}{\sigma_{\mathrm{ITD}}(c,\varphi)^2}} \qquad (4)$$

To avoid extremely high values due to variances close to zero a limit to the confidence was set. In a final step the confidence was normalized to the maximal confidence for all cues and all frequencies in order to have values in the range 0 to1. The resulting confidences are shown in Fig. 5.

D. Integration of the Cues

In a final step the different localization cues are integrated to form a localization estimate. For the integration an approach inspired by neural receptive fields was used. Details

[1] For the sake of simplicity only the ITD cue is shown, but an identical procedure was used for the remaining cues

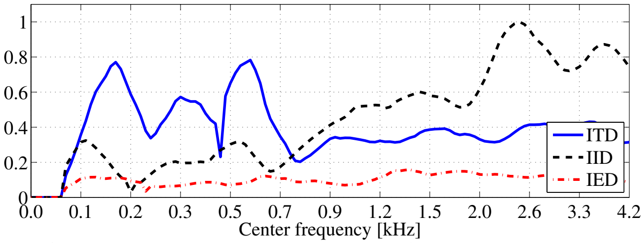

Fig. 5. The confidence values η for the different cues over the channel center frequency. To be complete also the confidence of the IED cue is shown but due to its low confidence it hardly contributes to the final result.

on this approach can be found in [1]. In the implementation used here a Gaussian function is centered for each angle and each frequency channel at the cue value found in the mapping matrix. The activation of the node

$$A_{\mathrm{ITD}}(c,\varphi,k) = w_{\mathrm{ITD}}(c,\varphi,k)\cdot \exp\left(-\frac{(\mathrm{ITD}(c,k)-\mathrm{ITD}_{\mathrm{M}}(c,\varphi))^2}{2\sigma(c)^2}\right) \quad (5)$$

represents how close the current measure $\mathrm{ITD}(c,k)$ at channel c and time instant k is to the cue value in the mapping matrix $\mathrm{ITD}_{\mathrm{M}}(c,\varphi)$ for the same channel and at angle φ. The parameter σ determines the width of the Gaussian kernel. The confidence weight $w_{\mathrm{ITD}}(c,\varphi,k) = \tilde{\eta}_{\mathrm{ITD}}(c)\cdot x(c,k)$ combines the previously calculated cue confidence $\tilde{\eta}_{\mathrm{ITD}}(c)$ and the energy of the underlying channel $x(c,k)$ after noise reduction ($x(c,k)$ is either $x_L(c,k)$ or $x_R(c,k)$ depending on which channel produced the maximum). The energy weighting enhances measures from signal parts with high energy as they normally are more reliable due to their better *Signal to Noise Ratio (SNR)*. Furthermore, a noise level dependent threshold $\delta_N(c)$ can be used for $x(c,k)$ so that only measurements where the energy of the underlying channel was above the noise level produce activations. For cases where the mapping is ambiguous, resp. non-injective, multiple activations for the same cue at different angles appear. This is a desired behavior as despite their ambiguity there is still information about the source location in these cue values. In an integration phase a histogram for the activations is build by summing over all channels K and cues:

$$H(\varphi,k) = \sum_{c=1}^{K} A_{\mathrm{ITD}}(c,\varphi,k) + A_{\mathrm{IID}}(c,\varphi,k) + A_{\mathrm{IED}}(c,\varphi,k) \quad (6)$$

In the histogram peaks form at the source location.

IV. Results

The performance of the system is illustrated by means of some results recorded with the dummy head mentioned in a conference room approximately of the size 7 m × 15 m and height of 3 m (reverberation time $\mathrm{RT}_{60} = 750\,\mathrm{ms}$). The walls are a large window front partly covered by blinds set open during the recording and wallpaper. On the floor was a carpet and the ceiling was normal wallpaper. Additionally an air conditioning and the PC fans were present in the room and adding to the noise floor. Though the results are only shown for this room we performed also tests with the real-time system in a smaller room. This room is approximately 4 m × 5 m and height of 2.5 m with two walls consisting of uncovered windows and two walls with normal wallpaper (reverberation time $\mathrm{RT}_{60} = 330\,\mathrm{ms}$). Also in this room the system performed good localization. The sampling rate was set to 48 kHz. We used a Gammatone filter bank with 128 channels where center frequencies are increasing logarithmically from 50 Hz to 5 kHz. Filter banks ranging up to 10 kHz or 15 kHz were also tested and yielded similar results. Before the envelope smoothing we applied a logarithm to the envelope signal. For the adjustment of $\tau = 180\,\mathrm{ms}$ we oriented ourselves at the estimated recovery time for humans from adaptation, the time constants in auditory models used as a front end for speech recognition, and the dynamics of the recorded signals[16]. The inhibition factor $\vartheta = 1.07$ was determined empirically. The noise threshold $\delta_N(c)$ was set to the noise floor.

A. Comparison to onsets

In Fig. 6 and 7 the results of our system are compared to a similar system using onsets and a fixed suppression window for following onsets instead of the maxima and the signal dependent inhibition proposed here. The upper plot of Fig. 6 shows the results of the onset based system for a speech signal presented via a loudspeaker at 1.3 m and 0° azimuth. In the lower plot the results for our maxima based

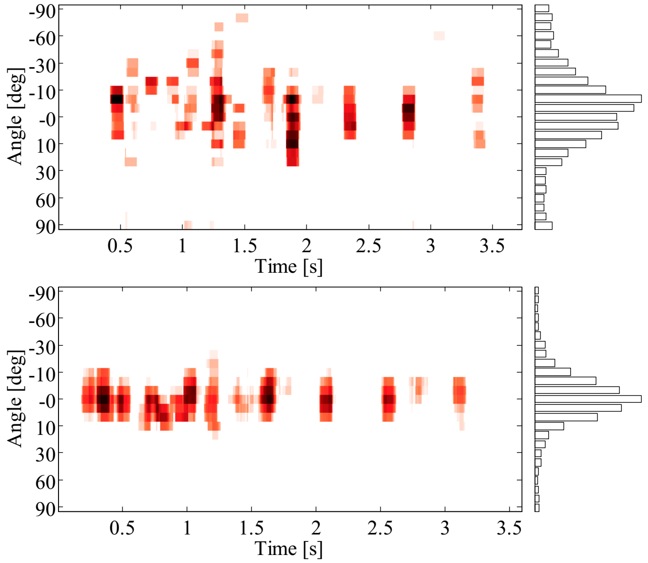

Fig. 6. Comparison between the use of onsets and maxima for cue evaluation. In the upper plot the localization results for one speech signal presented via loudspeaker at a distance of 1.3 m and 0° azimuth is shown. On the left hand side the histogram for each angle evaluated for each sample over all channels is shown. The right graph gives the sum of this histogram over all samples. The lower plot shows the results for the same signal when using the maxima.

system are shown. The left graph shows the activations of the different angles over time. A smoothing along the time was performed with a Gaussian window of $100\,\mathrm{ms}$ width. On the right graph a histogram for all the activations summed up over time is given. As can be seen the activations are much better concentrated on the real location in the case of our system compared to the onsets. Figure 7 shows the same comparison but this time with an additional speech signal presented via loudspeaker at 90° and a distance of $3\,\mathrm{m}$. Both signals were presented at the same loudness. The second source significantly deteriorates the localization of the first source in the case of the onset based system (compare the upper plot in Fig. 7). Several side maxima are present between the location of source 1 and 2 but no peak forms at the true location of source 2. In the case of the maxima based system the impairments in the localization of the source at 0° due to the additional source at 90° are much smaller and in the histogram summed over time the shape of the peak is only changed a little (compare the lower plot in Fig. 7). The second source at 90° does hardly appear in the graph and the summed histogram but this is largely due to the fact, that the second source is at $3\,\mathrm{m}$ compared to $1.3\,\mathrm{m}$ for the first source.

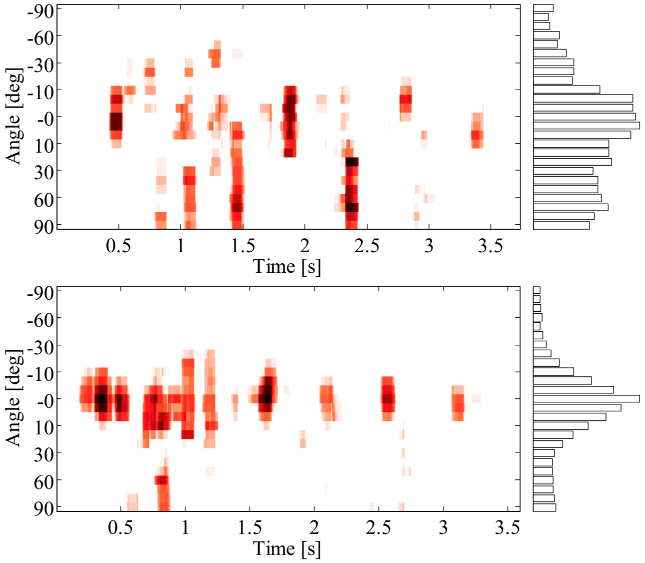

Fig. 7. Comparison between the use of onsets and maxima for cue evaluation. In the upper plot the localization results for two speech signals presented via loudspeaker. The first at a distance of $1.3\,\mathrm{m}$ and 0° azimuth the second at $3\,\mathrm{m}$ and 90° is shown. The lower plot shows the results for the same signal when using the maxima.

B. Localization in noisy conditions

In Fig. 8 localization results are given for a person talking at about $2\,\mathrm{m}$ distance and roughly 0° azimuth when additionally noise recorded from the fans of Asimo was presented via a loudspeaker directly from behind the head and music[2] from approximately $-80°$(from the left). The values for the distances and angles are only approximative as this was done with a real speaker standing in front of the system. For this reason we are also not able to give a precise SNR for the signals. As the signals started one after the other (the fan noise was already present before the recording started, a few seconds later the music set in and finally the speaker) we can give some approximative values though. We calculated these approximative values via the mean over the respective segments. The SNR between the music and the fan noise was about $-3\,\mathrm{dB}$ in the left ear and $-5\,\mathrm{dB}$ in the right ear. Higher SNR values in the left ear are due to the fact that the music was on the left side. The SNR of the speech signal to the combined music and fan noise was approximately $1\,\mathrm{dB}$ in the left ear and $2\,\mathrm{dB}$ in the right ear, differences in the ears are due to uncertainty of the true position and measurement errors. In the plot in Fig. 8 the music starts at $0\,\mathrm{s}$ and the speech signal at $2\,\mathrm{s}$. The part with only the fan noise present was cut out for visualization. As can be seen from the plot the fan noise and music are almost completely suppressed in the histogram by the noise reduction. The peak in the histogram on the right side is much wider than in the previous cases and the main peak is not at 0° but at 5°. It has to be taken into account that the absolute position of the speaker is not known and in the real time system the localization has a precision such that the head is facing the speaker after it turned to the speaker [1]. Fig. 9 shows a similar setup but this time the speaker was at roughly 60°. The music and the fan noise were kept at the same location and level. The SNR between the music and the fan noise was about $0\,\mathrm{dB}$, in the left ear

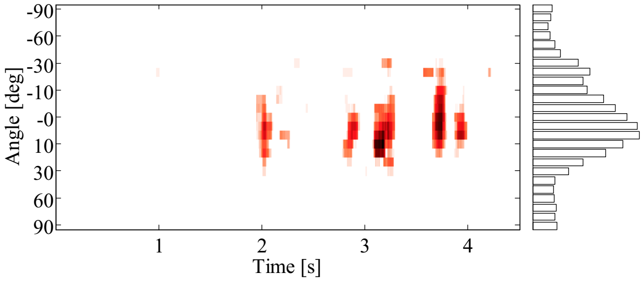

Fig. 8. Localization result for our system of a person talking at a distance of about $2\,\mathrm{m}$ and 0° when additionally fan noise from the back and music at $-80°$ were present.

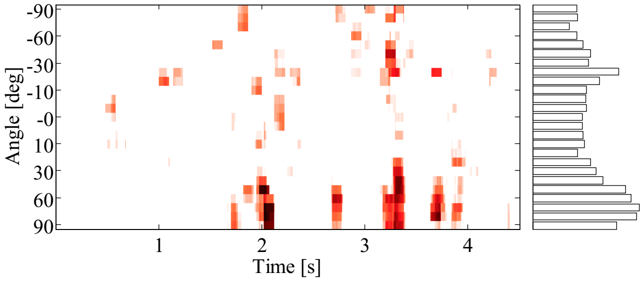

Fig. 9. Localization result for our system of a person talking at a distance of about $2\,\mathrm{m}$ and 60° when additionally fan noise from the back and music at $-80°$ were present.

[2] Wolfgang Amadeus Mozart Piano Sonata No. 11 in A major, K. 331 (Alla Turca), Allegretto

and $-2\,\mathrm{dB}$ in the right ear. Changes in the values compared to the previous setup are due to the imprecise measurement as the setup was not changed. As SNR between speech and combined fan and music we estimated in this scenario $-3\,\mathrm{dB}$ in the left ear and $0\,\mathrm{dB}$ in the right ear. The SNR also varies due to the fact that the speaker could not utter at exactly the same loudness in each trial. The music starts at $0\,\mathrm{s}$ and the speech signal at $1.8\,\mathrm{s}$. As can be seen the main peak forms at 70° and some side peaks in the direction of the music are present. In general we see a trend for more precise localization at around 0° and decreasing performance at the outer regions. This is due to the fact that the cue sensitivity is highest at 0° and decreases to the side. Finally Fig. 10 shows a setup where the speaker was at roughly -30°, hence at the same side as the music. The remaining setup remained unchanged. The SNR between the music and the fan noise was about $-2\,\mathrm{dB}$, in the left ear and $-4\,\mathrm{dB}$ in the right ear. As SNR between speech and combined fan and music we estimated $-1\,\mathrm{dB}$ in both ears. The music starts at $0\,\mathrm{s}$ and the speech signal at $1.9\,\mathrm{s}$. As can be seen the main peak forms at -30° and some side peaks in the direction of the music are present. The music interferes more with the localization in this case as it is on the same side but the speaker is still correctly localized.

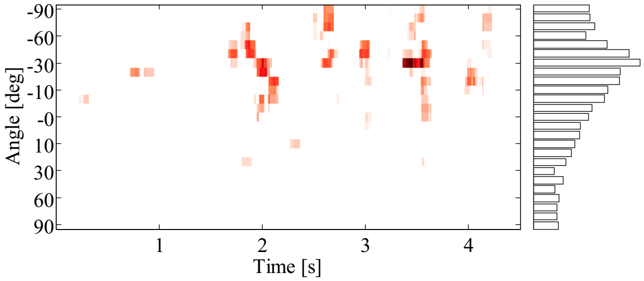

Fig. 10. Localization result for our system of a person talking at a distance of about $2\,\mathrm{m}$ and -30° when additionally fan noise from the back and music at -80° were present.

V. Discussion

We developed a system which is able to perform sound source localization with 2 microphones in strongly echoic and noisy conditions. Our system was inspired by the human auditory system which is reflected in the binaural setting with a dummy head, the auditory preprocessing by the Gammatone filter bank, the use of zero crossings, the neural integration of the cues, and the modeling of the precedence effect. Especially for the precedence effect we largely modified and extended previous approaches which relied on onsets. Our system uses the maxima of the envelope signal and performs a signal dependent, not fixed as in previous systems, inhibition of shortly following maxima. This inhibition is overwritten by a following stronger maximum. These properties are in line with the findings from psychoacoustics. We compared the results of our system to an onset based system in a single and two source scenario. There we could show that the localization results of our system are more reliable and precise than those of the onset based system. Furthermore we evaluated the performance of the system in a three source scenario with very bad SNR for the target signal. Here performance degrades in comparison to the single and two source scenario but results are still good enough for the use on a robot. The use of the zero crossings enabled the implementation of the system in real-time[1].

VI. Acknowledgments

We thank Mark Dunn for his support with technical problems and Julian Eggert for fruitful discussions.

References

[1] T. Rodemann, M. Heckmann, B. Schölling, F. Joublin, and C. Goerick, "Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping," in *Proc. of the Int. Conf. on Intelligent Robots & Systems (IROS)'06*. IEEE, 2006, p. submitted.

[2] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 34, no. 3, pp. 1526–1540, 2004.

[3] D. Bechler and K. Kroschel, "Reliability criteria evaluation for TDOA estimates in a variety of real environments," in *Proc. Int. Conf. Acoust. Speech and Sig. Proc. (ICASSP)'05*, Philadelphia, PA, 2005.

[4] E. Jan and J. L. Flanagan, "Sound source localization in reverberant environments using an outlier elimination algorithm," in *Proc. ICSLP '96*, vol. 3, Philadelphia, PA, 1996, pp. 1321–1324.

[5] K. Nakadai, H. Nakajima, K. Yamada, Y. Hasegawa, T. Nakamura, and H. Tsujino, "Sound source tracking with directivity pattern estimation using a 64 ch microphone array," in *Proc. Int. Conf. Intelligent Robots and Systems (IROS)'05*, Edmonton, Canada, 2005, pp. 196–202.

[6] J. Murray, S. Wermter, and H. Erwin, "Auditory robotic tracking of sound sources using hybrid cross-correlation and recurrent networks," in *Proc. Int. Conf. Intelligent Robots and Systems (IROS)'05*, Edmonton, Canada, 2005, pp. 891–896.

[7] H. Okuno and K. Nakadai, "Active audition for humanoid robots that can listen to three simultaneous talkers," *Journ. of the Acoust. Soc. of America (JASA)*, vol. 113, no. 4, p. 2230, 2003.

[8] E. Berglund and J. Sitte, "Sound source localisation through active audition," in *Proc. Int. Conf. Intelligent Robots & Systems (IROS)'05*, Edmonton, Canada, 2005, pp. 509–514.

[9] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Körner, "A probabilistic model for binaural sound localization," *to appear in IEEE Trans. on Systems, Man and Cybernetics - Part B*, 2006.

[10] T. Zahn, "Neural architecture for echo supression during sound source localization based on spiking neural cell models," PhD. Thesis, TU Ilmenau, Ilmenau, Germany, 2003.

[11] B. C. J. Moore, *An introduction to the psychology of hearing*, 5th ed. London: Academic Press, 2003.

[12] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank," Apple Computer Co., Tech. Rep., 1993, technical report #35.

[13] Y.-I. Kim, S. J. An, R. M. Kil, and H.-M. Park, "Sound segregation based on binaural zero-crossings," in *Proc. Int. Conf. on Spoken Lang. Proc. (ICSLP)'05*, Lisbon, Portugal, 2005, pp. 2325–2328.

[14] D. Kim, S. Lee, and R. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech and Audio Proc.*, vol. 7, no. 1, pp. 55–69, 1999.

[15] C. Kaernbach and L. Demany, "Psychophysical evidence against the autocorrelation theory of auditory temporal processing," *Journ. of the Acoust. Soc. of America (JASA)*, vol. 104, pp. 2298–2306, 1998.

[16] M. Holmberg, D. Gelbart, and W. Hemmert, "Automatic speech recognition with an adaptation model motivated by auditory processing," *IEEE Trans. Speech and Audio Proc.*, vol. 14, no. 1, pp. 43–49, 2006.