Real-time acoustic source localization in noisy environments for human-robot multimodal interaction

Vlad M. Trifa[1,2], Ansgar Koene[2,3], Jan Morén[2], Gordon Cheng[2,4]
[1]Institute for Pervasive Computing, ETH Zurich, 8092 Zurich, Switzerland
[2]ATR CNS Humanoid Robotics and Computational Neuroscience
[3]Knowledge Creating Communication Research Center, NICT
2-2-2 Hikaridai, "Keihanna Science City", Kyoto 619-0288, Japan
[4]JST-ICORP Computational Brain Project, 4-1-8 Honcho, Kawaguchi, Saitama, Japan
vlad.trifa@ieee.org, jan.moren@lucs.lu.se, {koene,gordon}@atr.jp

***Abstract*— Interaction between humans involves a plethora of sensory information, both in the form of explicit communication as well as more subtle unconsciously perceived signals. In order to enable natural human-robot interaction, robots will have to acquire the skills to detect and meaningfully integrate information from multiple modalities. In this article, we focus on sound localization in the context of a multi-sensory humanoid robot that combines audio and video information to yield natural and intuitive responses to human behavior, such as directed eye-head movements towards natural stimuli. We highlight four common sound source localization algorithms and compare their performance and advantages for real-time interaction. We also briefly introduce an integrated distributed control framework called DVC, where additional modalities such as speech recognition, visual tracking, or object recognition can easily be integrated. We further describe the way the sound localization module has been integrated in our humanoid robot, CB.**

I. Introduction

Human-robot interaction is a significant area of interest in robotics which has attracted a wide variety of studies in recent years. Most of these projects focused on particular aspects such as speech recognition [1], visual tracking and foveation [2], object recognition [3], learning through manipulation [4], or imitation learning [5] as isolated, independent problems rather than interconnected components of a highly integrated control architecture. Given that communication between humans is typically a multimodal process that simultaneously uses verbal and non-verbal cues to convey meaning, similar multimodal perceptual abilities would greatly enhance the ability of a robot to interact with humans. With the goal to attain flexible human-robot interaction, we propose a comparison of four common acoustic source localization algorithms. We discuss how robust human speaker localization in noisy environments can be achieved in real-time by using multimodal information fusion to drive gaze shifts and focus of attention of a humanoid robot on auditory events.

For sound source localization, the error in direction is about 8-10 degrees in humans [6], which is too coarse to separate sound streams from a mixture of sounds. In visual processing, there are other problems such as narrow visual field of an ordinary camera or visual occlusions. Thus each modality has its own weaknesses, but they can be overcome by integrating the visual and auditory information to take advantage of the best of both worlds [7]. Additionally, integration of visual information in speech recognition (for example lip-reading) can greatly improve the performance of speech understanding (see [8] for an extensive overview on this topic).

Recently improved human-computer interfaces combining several sensory modalities have started to appear [9]–[14]. Unfortunately, most of these multimodal systems are based on post-perceptual integration, where modalities are treated as individual systems and only their output is merged in the final step making them too rigid to deal with complex human behavior. In humans and animals, behavioral and physiological evidence suggests that information merging happens also at earlier levels and that unimodal sensory information processing can be strongly biased by other sensory modalities [15]. The response enhancement during multi-sensory activation is supported by recent functional imaging (fMRI) [16]. This close link between perceptual systems may well be a key to human-like perceptual flexibility.

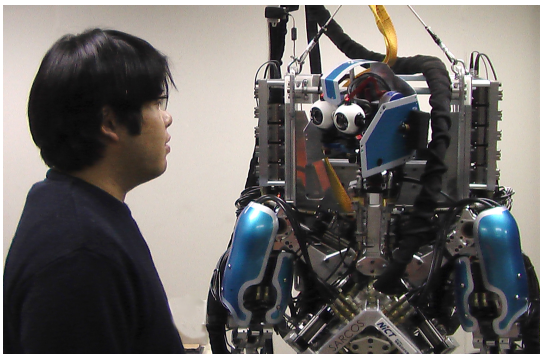

Fig. 1. The humanoid robot CB used in our experiments, orients his head to face human speaker based on acoustic signals. Copyright ATR. The robot was developed by SARCOS for NiCT.

Our research investigates efficient ways to combine acoustic and visual information for robust interaction by using biologically inspired models of attention shift and motor control. However, this paper will focus only on the sound localization problem and what properties are desired for multimodal integration (the actual integration mechanism is detailed in [17]). This will enable speech recognition systems to separate talkers from each other and from environmental sounds. We will briefly present the general architecture of the perceptual system we are developing for the Humanoid Robot CB [18] (shown in Fig. 1), featuring sensory (audio-visual) integration,

bottom-up salience detection, top-down attentional feature modulation, and reflexive gaze shifting.

II. Sound localization for multimodal interaction

In this article, the term localization refers only to the estimation of the Direction of Arrival α (DOA) of a sound, and not the actual position of the acoustic source. Jeffress proposed in his so-called Duplex Theory [19] that two primary cues are mainly used for sound localization – *Interaural Time Difference* (ITD) and *Interaural Level Difference* (ILD) – and their combination can lead to robust DOA estimation using the whole audible frequency spectrum. For a sound wave that strikes a spherical head of radius r from a direction specified by the azimuth angle α, the difference in the length of the straight-line path to the two ears is $2r\sin(\alpha)$, which corresponds to a time difference of $2r\sin(\alpha)/c$, where c is the speed of sound (approx. 334 m/s).

At low frequencies the wavelength of the sound is much larger than the head diameter, so the phase difference between the signals can be estimated with no ambiguity, but for high frequencies there can be several cycles of shift, leading to ambiguity for the ITD. This ambiguity can be resolved by using the ILD. Incident sound waves are diffracted by the head, resulting in a significant difference in the sound pressure on the two ears – the ILD, which is highly frequency dependent. At low frequencies (below about 1.5 kHz), there is hardly any difference in sound pressure at the two ears, but at high frequencies above about 1.5 kHz (where there is ambiguity in the ITD measure) the difference in pressure is sufficient to lateralize a sound signal.

Unfortunately, the relationship between a source signal and the pressure developed at the ear drums is not only difficult to model analytically because of the complexity of the head geometry and range of wavelengths to consider, but also varies according to the azimuth, elevation, range of the source, and the environment itself. These effects can be captured in the so-called *Head-Related Transfer Function* (HRTF), which can be used to localize sound sources in 3D with only two ears. So far, most of our knowledge of HRTFs has come from direct experimental measurements in anechoic room with microphones inserted in the ears of subjects. Several approaches where HRTF are used for localization have been proposed (e.g., [13]), but these methods are not efficient in real-world environments. A model of visuo-motor learning based on visual feedback in the barn owl has been proposed in [20], but uses only a single camera and localization is limited to the horizontal plane. A more effective model that learns sensory-motor coordination of auditory-evoked reflexes is proposed in [21], however the focus in the present paper is on simple auditory perception methods suited for multimodal processing, rather than elaborate models of self-calibration and motor response learning. An extensive review of models for auditory perception based on time delay estimation is given in [6]. Recent interesting results on biologically plausible models of auditory perception using biological spiking neural networks are proposed in [22], [23].

When choosing an appropriate auditory sound localization method however one needs to consider not only performance and accuracy, but also the computational and hardware cost. In a multimodal system, such as a humanoid robot, this cost-benefit analysis must include other modalities, such as vision, which are inherently better suited for fine localization performance. Furthermore, active audio and visual perception in robots should involve dynamic exploratory reorientation of sensors to improve the quality of the perceived. To achieve real-time performance and have a responsive and flexible system that can quickly tune to changes in the environment, we will follow the general organization of an integrated control architecture proposed in [7]. Within this context, very high resolution sound localization may therefore be a waste of resources.

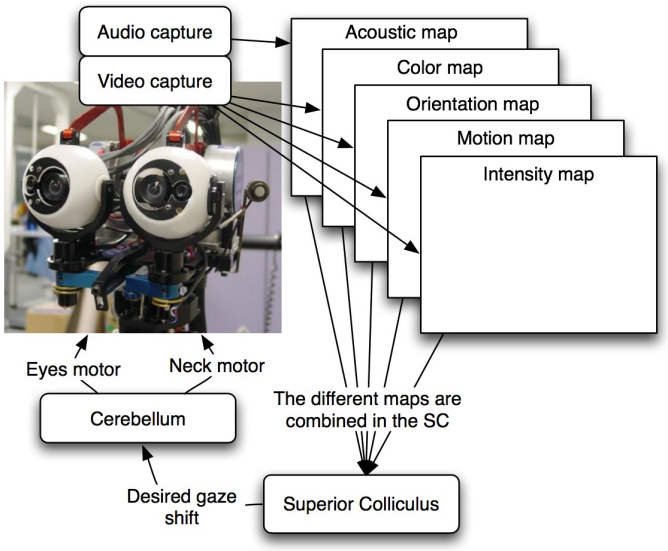

Fig. 2. The global architecture of our model. Different spatial maps for every modality are combined in the *Superior Colliculus* model, with possible high-level top-bottom biasing influence. The Superior Colliculus model computes a desired gaze shift that is transmitted to the Cerebellum model, where the gaze shift is transformed into the actual low-level motor commands for both eyes and neck movement.

III. Methods

Following a brief description of the humanoid robot CB and its multimodal perception architecture, we present four models we have implemented in Matlab and compare their performance on the same stereo input signal recorded with the robot.

A. Humanoid robot CB

CB stands for *Computational Brain*, and refers to an elaborate humanoid robot developed by SARCOS [18]. CB possesses 50 degrees of freedom (DOF), and the head has 7 DOF (eyes pan and tilt independently, and roll/pitch/yaw for the neck) and is similar in performance with a human head.

The microphones are located on the head of our humanoid robot as illustrated in Fig. 2, positioned 17.6 cm apart. Supercardioid microphones Shure MX184BP, pre-amplified by a STICK-ON STM-2 low-noise pre-amp with variable gain, have been used for these experiments. The signals used in the Matlab analysis were recorded directly from the robot microphones in 16 bits wave format. The microphones were connected to the line-in input port of a sound card on a desktop computer, where sampling was done using the standard Microsoft API in a software module which is a component of our complete control system, as described in the next section.

The different sensory inputs are decomposed into several parallel streams, each corresponding to one type of visual

or auditory feature. Similar to the bottom-up visual salience model from Itti and Koch [24], these features are used to generate spatial maps that encode areas of interest. These maps are combined to generate a global saliency map that emphasizes locations that stand out from their surrounding. The global saliency map is used as an input to a winner-take-all neural network which is used to compute the most salient area. While this approach is purely bottom-up, top-down effects can be introduced by biasing the weights when combining the conspicuity maps or by introducing lateral inhibition when computing local feature maps. In order to achieve real-time operation of the complete visual and acoustic attention system a distributed architecture is essential. For this reason, the whole system has been implemented on a cluster of heterogeneous computers using the DVC framework [25], [26].

B. Signals preprocessing

The recorded signals are filtered using a band-pass filter in the range of human hearing (from pilot experiments we found that good results are obtained when using as lower cutoff $f^1_{pass} = 1000$ Hz, and upper cutoff $f^2_{pass} = 2000$ Hz) to remove unwanted noise. Blocks of 256 samples are continuously fetched from the hardware and a simple sound detection algorithm is used to detect when acoustic activity takes place. Initially, a statistical model of the power spectrum E in a limited frequency range of background noise is estimated (assuming E follows a normal distribution $N(\mu, \sigma^2)$), and the energy in this frequency range is monitored until it reaches a detection threshold ($E \geq \mu + \beta\sigma, \beta = 3$). Once the threshold is reached, the detected block of audio data is fed to the actual localization module which is described in the following section.

C. Acoustic localization

A multitude of methods exist for accurate acoustic source localization, some of them using very elaborate models of the environment or of the signals. The interested reader is invited to consult [27] for a systematic overview of the state-of-the-art of time-delay estimation techniques. For our purpose, however, complex optimal localization techniques are too computationally expensive. We therefore compare simple localization models and identify properties required to obtain robust human-robot interaction.

1) Generalized Cross-correlation (GCC): The most straightforward method to estimate the time-delay between two signals x_l and x_r is *the cross-correlation* (CC), which consists of summing the signals for every possible delay between two microphones, and then select the delay for which the sum is maximal as an estimate of the time delay. This procedure is very sensitive to noise and reverberations, thus should be discarded for localization in real environments. A more efficient and generic method to estimate the time delay is the *Generalized Cross-correlation* (GCC) that is defined as follows:

$$R_{l,r}(\tau) = \int_{-\infty}^{\infty} X_l(\omega) X_r^*(\omega) \exp^{j\omega\tau} d\omega$$

where X is the Fourier transform of the signal x, and X^* denotes the conjugate of the Fourier transform. The GCC is a more robust method than direct CC that is based on pre-filtering the input signals so as to take into account that a finite time window of observation is used in reality, and that multiple sound sources or echoes may be present [20]. The problem with this simple method, is that correlation between samples is usually large so peaks can be quite wide, resulting in lower precision in the final result.

2) GCC with Phase Transform (PHAT): The manifestation of interfering signals is easier to detect in the frequency domain, and an alternative approach for DOA estimation considers the signals x_l and x_r in the frequency domain to remove signal interference in the real life situations. To extract DOA from frequency domain using GCC, we compute the inverse Fourier Transform of the signal cross-power spectrum scaled by a weighting function. One instance of GCC weighting function is the Cross-power Spectrum Phase, also called the *Phase Transform* (PHAT). This weighting places equal importance on each frequency band. By dividing the spectrum by its magnitude it de-emphasizes portions of the spectrum that are suspected to be corrupt. This process results in a constant energy concentrated over all frequencies so that the correct DOA can be found by high coherence between the two signals. Experiments have shown PHAT to be very robust to noise and reverberation. PHAT is defined as follows

$$R_{l,r}(\tau) = \int_{-\infty}^{\infty} G(\omega) X_l(\omega) X_r^*(\omega) \exp^{j\omega\tau} d\omega$$

where $G(\omega)$ is a weighting factor as described earlier, and in the case of PHAT weighting it is computed as follows:

$$G^{PHAT}(\omega) = \frac{1}{|X_l(\omega) X_r^*(\omega)|}$$

Other weighting factors can also be used and are presented in [28]. The problem with PHAT is when interferences are dominant over the signal (resulting in a low signal-to-noise ratio), PHAT results for DOA estimation will be unreliable. Due to room reverberations and environmental noise a number of undesired local maxima can be found in the GCC function. To emphasize the GCC value at the true DOA value different weighting functions have been investigated. For real environments the Phase Transform (PHAT) technique has shown best performance and is applied by most DOA estimation algorithms. Given that we are mainly interested in localizing speech, one should aim to detect voice in the spectrum, and the weight function should emphasize regions that are likely to contain voice components. The idea behind this technique is that when no single frequency dominates, the effects of reverberation cancel out when averaged over many frequencies.

3) Moddemeijer Information theoretic delay criterion (MODD): The main weakness of frequency-domain implementations of delay estimators (e.g., GCC) lies in the difficulty to correctly estimate the spectrum of short signals. To counter this problem, Moddemeijer [29] developed a time-domain implementation of an advanced delay estimation that uses information theory to maximize the probability of source location using mutual information measures between binaural signals. A detailed mathematical explanation of this method can be found in [23]. The main advantage of the method is the proven unambiguity of the criterion.

4) Cochlear filtering (COCH): An alternative method for time delay estimation is to use a model of cochlear filtering on the input signals before a cross-correlation, and we used a modified version of the cross-correlation model found in

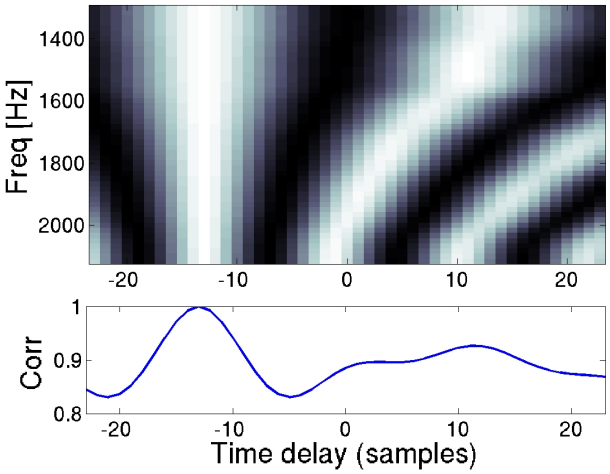

Fig. 3. Correlation of a simulated signal using a gammatone filter-bank. White areas represent high activity and correlation. One can see the number of side lobes increase with the frequency. *Top*: Correlation for different delays and frequencies. *Bottom*: Value of the activity summed over all frequencies.

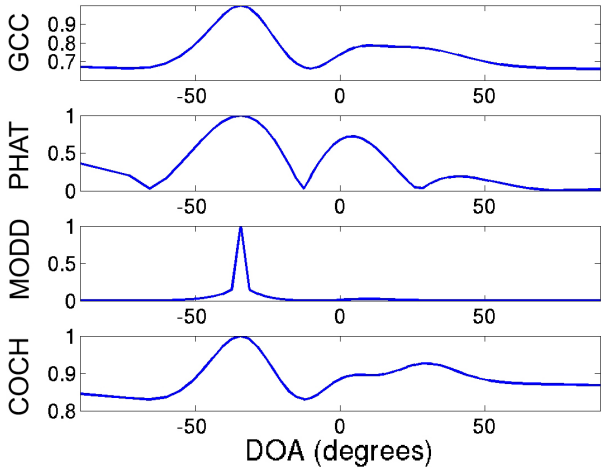

Fig. 4. Results of the DOA estimation with simulated signals for a spoken sentence located at -30°. *From top to bottom:* cross correlation (GCC), generalized cross correlation with phase transform weighting (PHAT), Moddemeijer information theoretical approach (MODD), and cochlear filtering (COCH).

Akeroyd's Binaural Toolbox [30]. At first, the signal of each microphone is filtered using a gamma-tone filter-bank (we used ERB filter-bank from [31]) that decomposes the signal into several streams having different frequency sensibilities, similar to the frequency transformation performed by the cochlea. For each of these streams a cross-correlation is performed resulting in a bi-dimensional coincidence detection map that shows the correlation of the binaural signals for different frequency ranges and time delays (see Figure 3). In theory the performance should be superior to the other approaches discussed in this paper as high coherence for a specific delay should be visible in the coincidence detector - at least for wide-band signals. The proposed filter-bank, based upon Lyon's cochlear model [32], is composed of 50 gamma-tone filters distributed between 700 Hz and 2000 Hz.

IV. Experiments

We performed several tests to compare the results obtained by these four common approaches to binaural time delay estimation with various noise conditions. The present study was conducted using high quality sound samples extracted from the RWCP Sound Scene Database in Real Acoustic Environment. Four sentences - two in Japanese, and two in English - by different speakers were used (duration between 3 and 5 seconds). These samples were played through a loudspeaker located at -30° in front of the robot. The resulting sound stimuli were recorded using the microphones mounted on the head of the robot. The robot is located in a room with a large amount of background noise due to cooling fans. We ran 3 series of tests with various amount/types of background noise to analyze how the sound localization methods perform in case of noisy situations. For the three experiments the background was composed of, no added noise (only room noise at 62.9dB), added electronic music (total noise of 76dB), and added white noise (total noise of 79dB), respectively.

A. Simulation

To ensure the correctness of our methods, we compared GCC, PHAT, MODD, and COCH in a simulation where the binaural signal is obtained by copying the signal from one microphone to the channel of the other microphone and shifting it by a fixed number of samples. For all our tests and methods, the fixed delay was consistently estimated correctly. This result, however can not be extrapolated to real-world situations for two reasons 1) the signals in both channels are perfectly identical, which is hardly the case in a real experiment, and 2) an artificial delay assumes that all the background noise and reverberations emanate from the same direction as the signal itself, thus do not affect the DOA estimation procedure. Time delay estimation of a simulated (artificial delay = -12 samples) binaural signal can be seen in Figure 4.

It is interesting to point out that even in simulation, secondary lobes can be seen in the correlation graph, especially for GCC and PHAT (see Figure 4), which can lead to ambiguity in delay estimation. However, MODD gives a single thin peak, suggesting a higher spatial discrimination resolution than the other approaches.

B. Robot implementation

The input data was based on the same recordings as in the simulation experiment, but the actual left and right channels were used instead of replacing one of the channels with a shifted version of the signal from the other channel. In these experiments, the performance was significantly lower than in the simulation, probably due to the high amount of noise in the experiment room.

During our experiments, the background noise was coming from the left side of the robot due to a gigantic compressor. When the noise was played from the speaker on the left (from the same direction as the background noise), localization was close to the exact direction for all methods in all noise conditions (CCF: $\mu = -35.1°, \sigma = 3.7°$; PHAT: $\mu = -33.4°, \sigma = 1.2°$; MODD:$\mu = -33.8°, \sigma = 0.6°$; COCH: $\mu = -36.7°, \sigma = 2.6°$), and performance was even most stable in the white noise added case. From our results we noticed that the results in the generalized cross-correlation had the highest variation, even for the same noise condition. PHAT had the closest prediction to the actual angle, while MODD was less close but was the most stable algorithm.

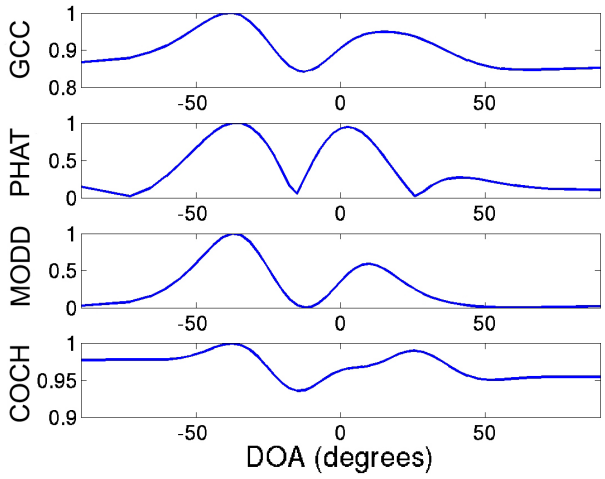

Fig. 5. Results of the DOA estimation with real signals, for a spoken sentence located at -30°. *From top to bottom:* cross correlation (GCC), generalized cross correlation with phase transform weighting (PHAT), Moddemeijer information theoretical approach (MODD), and cochlear filtering (COCH). All four methods yield the correct angle, but COCH and PHAT have quite large secondary lobes that can lead to ambiguous results. On the other hand, MODD performs the best, having the smallest secondary lobe and the thinnest main peak.

In the cochlear filtering method, coincidence detectors are run over several frequencies to improve the chances of correct localization. If the signal is wide band and the signal-to-noise ratio is greater than one (SNR $\geq$ 1), then this approach works very well, and is able to localize the sound accurately. However, the method becomes unstable when the signal is narrow band, while the noise is wide band (as is the case with speech). In the absence of a top-down attention system that dynamically selects which frequency bands to attend to, the cochlear model suffers from an internal conflict between signal localization (from the frequency bands that contain signal information) and noise localization (from the other bands).

Since most human speech signals are located between 1-3 kHz, our tests have shown that systematic and accurate localization is obtained in low noise environments, when a band-pass filter is used on the signals that filter out signals not located within this frequency band.

V. Discussion

We have implemented and compared the performance of four common algorithms to determine the Direction Of Arrival (DOA) of sound signals in noisy environments.

From our results we noticed that the generalized cross-correlation (GCC) had the highest variation, GCC with Phase Transform (PHAT) had the greatest accuracy, while Moddemeijer Information theoretic delay criterion (MODD) was most reliably precise. The Cochlear filtering (COCH) was handicapped by the lack of dynamic top-down frequency band selection and consequently did not perform quite as well as the other methods.

PHAT performs extremely well in reverberant environments, but is very sensitive to noise, given that the spectrum is flattened, therefore is not a very appropriate candidate for our purpose. In spite of MODD's high reliability, this method may not be the most desirable for humanoid robots because it can not easily integrate top-down modulation. MODD does not deal with the frequency domain, nor does it use biologically plausible mechanisms to infer source location. Given that our work deals with attention, it is desirable to have a simple mechanism to introduce top-down selectivity in both audio and visual perception. This is one of the strengths of a cochlear filter model in which frequency bands can be differently modulated according to interest [33]. Attention based frequency-band modulation is in fact required for good DOA performance in wide band noise with the COCH model.

When audio signals are the only information that can be used to track an acoustic source, efficient methods that rely on spatially distributed microphone arrays (e.g., [34]) are needed to achieve accurate results. However, such applications require dedicated hardware, optimized for acoustic localization, and due to the high computational costs associated with such complex algorithms, real-time processing can hardly be obtained. For human-robot interaction, where robustness in performance can be obtained through multimodal integration, it is preferable to use only two microphones associated with simple localization techniques.

All frequency domain based Interaural Time Delay (ITD) methods produce secondary peaks that can confuse the DOA estimation. These secondary peaks are caused by the aliasing that occurs for signals with a frequency higher than 1.5 kHz. As suggested by Jeffress [19], Interaural Level Differences ILD could be used as an initial coarse estimate of the spatial areas the signals could be in, and this information can be used to emphasize the activity in the spatial map at the locations where the source is more likely to be present. Unfortunately the current design of the CB robot head is not well suited for this.

Extensive studies of the sound localization apparatus of the barn owl [35] revealed the mechanism that permit these animals to locate sounds with great accuracy. In addition to the learning of orientation behavior and accurate tuning of motor response, the structural properties of the head and facial feather were found to be of vital importance. This suggest that two ears, augmented with pinnae-like reflectors, can be sufficient for robust sound localization, and that one should use a similar physical model when it comes to designing acoustic processing models.

One might be tempted to add winner-take all (WTA) processing to the output cross-correlator in order to amplify the highest peaks and level out all secondary peaks, as has been done in [23]. When using an audio-visual system, such as a humanoid robot, this should be avoided since this would hinder the cross-modal facilitation of stimulus localization. In a multi-modal system each separate modality provides only part of the total percept. The initial feed-forward processing should therefore retain as much information as possible to allow for percept changes due to multi-modal information correlation.

When using audio-visual response enhancement, one can significantly improve the probability of locating the most salient location in the environment. In addition, a flexible fusion scheme can ensure rapid gaze shift even towards locations not in the visual scene, followed by visual re-foveation. We are now in the process of implementing such multi-modal interaction to produce reflex-gaze shifts towards multi-modal stimuli. The Superior Colliculus module in our robot receives two basic types of inputs, bottom-up excitatory inputs from the sensory processing modules (currently consisting of the visual and auditory systems) and top-down inputs (inhibitory & excitatory) from higher cognitive processing modules. The

inhibitory top-down inputs mediate spatial attention and general sensitivity while the excitatory inputs drive deliberate cognitive controlled top-down gaze shifts. The output of the SC module is a desired gaze shift signal.

In complex systems such as humanoid robots one of the reasons for requiring top-down modulation of auditory sensitivity is motor noise. Motor noise is complex and often irregular because of the quantity and nature of actuators that may be involved in the head and body movement. A filtering scheme that dynamically suppresses motor noise is required so that the audition system can perform in real-time, with high responsiveness, avoiding the need for a "stop-perceive-act" cycle. The addition of an extra motor-noise recording system however requires additional hardware and computational complexity. Our next target is to integrate in our top-down selection a scheme that de-emphasize the frequency bands where motor noise is likely to be present.

In this paper, we highlighted four common sound source localization algorithms and compared their advantages for real-time interaction in the context of a multi-sensory humanoid robot. From our results we noticed that the generalized cross-correlation (GCC) had the highest variation, GCC with Phase Transform (PHAT) had the greatest accuracy, while Moddemeijer Information theoretic delay criterion (MODD) was most reliably precise. The Cochlear filtering (COCH) was handicapped by the lack of dynamic top-down frequency band selection and consequently did not perform quite as well as the other methods though performance was still adequate under reasonable SNRs. When considering integration with signals from other modalities and attention mechanism, COCH should be considered as a primary candidate since this method is most suited for top-down modulation.

ACKNOWLEDGMENTS

The authors would like to thank the support of the Keihanna branch of the National Institute of Communication Telecommunication (NiCT), Japan. The robot was developed for NiCT in a NiCT-ATR collaboration. In addition, the authors wish to acknowledge the continuous supports of members of SARCOS Research Corporation. Thanks also to Kai Welke, for his latest re-implementation of DVC.

REFERENCES

[1] R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, and V. Zue, "Survey of the state of the art in human language technology," 1997.

[2] A. Ude, C. Gaskett, and G. Cheng, "Foveated vision systems with two cameras per eye," in *Proc. IEEE Int. Conf. Robotics and Automation*, Orlando, Florida,, May 2006, pp. 3457–3462.

[3] K. Welke, E. Oztop, A. Ude, R. Dillmann, and G. Cheng, "Learning feature representations for an object recognition system," in *IEEE-RAS/RSJ International Conference on Humanoid Robots (Humanoids 2006)*, December 2006.

[4] P. Fitzpatrick, "From first contact to close encounters: A developmentally deep perceptual system for a humanoid robot." Ph.D. dissertation, Massachussets Institute of Technology, 2003.

[5] S. Schaal, A. Ijspeert, and A. Billard, "Computational approaches to motor learning by imitation," In Frith, C. D.;Wolpert, D. (eds.), The Neuroscience of Social Interaction, Oxford University Press, pp. 199–218, 2005.

[6] J. Blauert, *Spatial Hearing : The psyhcophysics of Human Sound Localization*. MIT Press, 1996.

[7] G. Cheng, A. Nagakubo, and Y. Kuniyoshi, "Continuous humanoid interaction: An integrated perspective - gaining adaptivity, redundancy, flexibility - in one," *Robotics and Autonomous Systems*, vol. 37, pp. 161–183, 2001.

[8] C. Benoit, J.-C. Martin, C. Pelachaud, L. Schomaker, and B. Suhm, *Audio-visual and Multimodal Speech Systems*. In D. Gibbon (Ed.) Handbook of Standards and Resources for Spoken Language Systems - Supplement Volume, to appear., 1998.

[9] R. Bischoff and V. Graefe, "Integrating vision, touch and natural language in the control of a situation-oriented behavior-based humanoid robot," in *IEEE International Conference on Systems, Man, and Cybernetics*, 1999.

[10] R. Brooks, C. Breazeal, M. Marjanovic, B. Scassellati, and M. Williamson, "The cog project: Building a humanoid robot," in *Computation for Metaphors, Analogy and Agents, Springer Lecture Notes in Artificial Intelligence*, C. Nehaniv, Ed., vol. 1562, 1998.

[11] P. McGuire, J. Fritsch, J. J. Steil, F. Roethling, G. A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter, "Multi-modal human-machine communication for instructing robot grasping tasks," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Lausanne, Switzerland*, 2002, pp. 1082–1089.

[12] M. Shiomi, T. Kanda, N. Miralles, T. Miyashita, I. Fasel, J. Movellan, and H. Ishiguro, "Face-to-face interactive humanoid robot," in *Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004).*, 2004.

[13] Y. Matsusaka, T. Tojo, S. Kuota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi, "Multi-person conversation via multi-modal interface – a robot who communicates with multi-user," in *Proceeding of Eurospeech*, 1999, pp. 1723–1726.

[14] K. Nakadai, K. Hidai, H. G. Okuno, and H. Kitano, "Real-time active human tracking by hierarchical integration of audition and vision," in *Proceedings of 2001 IEEE-RAS International Conference on Humanoid Robots*, 2001.

[15] S. Shimojo and L. Shams, "Sensory modalities are not separate modalities: plasticity and interactions," *Current opinion in Neurobiology*, vol. 11, pp. 505–509, 2001.

[16] E. Macaluso, "Multisensory processing in sensory-specific cortical areas," *Neuroscientist*, vol. 12, no. 4, pp. 327–338, 2006.

[17] A. Koene, J. Moren, V. Trifa, and G. Cheng, "Gaze shift reflex in a humanoid active vision system," in *International Computer Vision Systems (ICVS 2007).*, 2007.

[18] G. Cheng, S.-H. Hyon, J. Morimoto, A. Ude, and S. C. Jacobsen, "Cb: A humanoid research platform for exploring neuroscience," in *IEEE-RAS/RSJ International Conference on Humanoid Robots (Humanoids 2006)*, 2006.

[19] L. Jeffress, "A place theory of sound localization," *J. Comp. Physiol. Psychol.*, vol. 41, pp. 35–39, 1948.

[20] M. Rucci, J. Wray, and G. Edelman, "Robust localization of auditory and visual targets in a robotic barn owl," *Journal of Robotics and Autonomous Systems*, vol. 30, no. 1-2, pp. 181–194, 2000.

[21] L. Natale, G. Metta, and G. Sandini, "Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head," *Robotics and Autonomous Systems*, vol. 39, no. 2, pp. 87–106, 2002.

[22] C. Schauer and P. Paschke, "A spike-based model of binaural sound localization." *International Journal of Neural Systems*, vol. 9, no. 5, pp. 447–452, 1999.

[23] C. Schauer, "Modellierung primärer multisensorischer mechanismen der räumlichen wahrnehmung," Ph.D. dissertation, Fakultät für Informatik und Automatisierung der Technischen Universität Ilmenau, 2006.

[24] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, no. 10-12, pp. 1489–1506, 2002.

[25] A. Ude, V. Wyart, L.-H. Lin, and G. Cheng, "Distributed visual attention on a humanoid robot," in *Proceedings of 2005 IEEE-RAS International Conference on Humanoid Robots*, 2005.

[26] G. Cheng, K. Welke, L.-H. Lin, P. Azad, and A. Ude, "A concurrent architecture for humanoid robots to emulate biological processes," *Robotics and Autonomous Systems*, 2007.

[27] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: an overwiev," *EURASIP Journal on applied signal processing*, vol. 2006, pp. 1–19, 2006.

[28] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1976.

[29] R. Moddemeijer, "An information theoretical delay estimator," Univ. of Twente, Enschende (NL), Tech. Rep. Report: 080.87.45, 1987.

[30] M. A. Akeroyd, "Binaural toolbox." [Online]. Available: http://www.biols.susx.ac.uk/Home/Michael_Akeroyd/

[31] M. Slaney, "Auditory toolbox," Interval Research Corporation, Tech. Rep. 1998-010, 1998.

[32] M. Slaney and F. R. Lyon, "On the importance of time - a temporal representation of sound," *Visual representations of speech signals*, 1993.

[33] S. N. Wrigley and G. J. Brown, "A model of auditory attention," University of Sheffield, Tech. Rep., 2000.

[34] C. Chen, A. Ali, H. Wang, S. Asgari, H. Park, R. Hudson, K. Yao, and C. E. Taylor, "Design and testing of robust acoustic arrays for localization and beamforming," in *IEEE IPSN*, 2006.

[35] K. E.I. and M. Konishi, "A neural map of auditory space in the owl," *Science*, vol. 200, no. 4343, pp. 795–797, 1978.