The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Pitch Extraction in Human-Robot Interaction

Martin Heckmann, Frank Joublin, Kazuhiro Nakadai

Abstract— **We present a system for real-time fundamental frequency, i. e. pitch, extraction on a humanoid robot. The system extracts pitch using an 8 channel microphone array mounted on the Honda humanoid robot in a realistic Human-Robot interaction scenario. The main building blocks of the system are a multi-channel signal enhancement followed by robust pitch extraction and tracking. The signal enhancement is based on 8 channel Geometric Source Separation. For the pitch extraction the signal is first transformed with a Gammatone filter bank into the frequency domain. Next a histogram of zero crossing distances is calculated from all filter bank signals. During the calculation of the histogram spurious side peaks resulting from harmonics and sub-harmonics of the true fundamental frequency are inhibited. The resulting histogram then serves as input to a grid based Bayesian tracker which deploys Bayesian filtering in a forward step and Bayesian smoothing in a backward step on a 100 ms time window. We demonstrate the performance of the system in a scenario where male and female speakers utter different phrases while standing at a normal interaction distance to the robot. For the evaluation we compare the pitch tracking results once obtained from a clean headset signal and once from the signals obtained from the robot. The results show that the tracking performance only degrades to a small extent in the realistic interaction scenario compared to the headset recordings.**

I. Introduction

Robots able to operate in the real world and to interact with humans require advanced capabilities to perceive their acoustic environment. As many of the sounds surrounding us are harmonic the fundamental frequency is an important cue to describe them. In the psychophysical literature the percept of fundamental frequency is called pitch. Especially for speech the pitch has a high importance as it distinguishes different intonation patterns, e. g. questions from statements, and in tonal languages also different words.

In human robot interaction the acoustic environments are in most cases very unfavorable. For a natural interaction the microphones capturing the sound signals are mounted on the robot and hence do not only capture the speech signal or other desired sounds but also a lot of background noise. This includes also the noise generated by the robot itself. Additionally, due to the long distance between the speaker and the robot, the influence of the room reverberations on the speech signal is also considerable. As a consequence the signals usually have low Signal to Noise Ratios (SNRs) and the task of extracting the fundamental frequency in such an environment is rather difficult.

M. Heckmann and F. Joublin are with the Honda Research Insitute Europe GmbH, D-63073 Offenbach/Main, Germany martin.heckmann@honda-ri.de, frank.joublin@honda-ri.de

K. Nakadai is with the Honda Research Insitute Japan Co. Ltd., Wako-shi, Saitama 351-0188, Japan nakadai@jp.honda-ri.com

Motivated by the astonishing performance of humans in such situations we took inspirations from models of pitch perception in the design of our pitch extraction algorithm [1]. In general the perceptual models fall into two categories [2]: Rate and place models.

The term rate in the context of pitch is reminiscent of the phase locking of the neurons in the auditory system. They fire, i. e. they produce a spike, at the same position in each cycle of the signal. Consequently they code the frequency of the signal.

Place refers in this context to the way the basilar membrane in the inner ear decomposes a signal into its frequency components. The place of maximal excitation of the basilar membrane moves from the basal end for very high frequencies to the apical end for very low frequencies. Thereby the frequency content of the signal is coded in the excitation pattern of the basilar membrane.

In a rate model the periodic structure of a harmonic signal is exploited. This can be done via application of an autocorrelation directly on the time signal [3], [4]. An extension of this idea is to perform the autocorrelation at the output of the basilar membrane and to sum up all individual results after some additional non-linearity [5]. The basilar membrane is commonly modeled via a band pass filter bank with bandwidths and center frequencies adapted to psychological data (also referred to as *Gammatone filter bank* [6]). Place or pattern matching models make use of equidistant lines in the spectrum of purely harmonic signals. The most prominent of these models is the one proposed in [7]. Here a comb filter with teeth at the fundamental frequency hypothesis and its harmonics is set up and applied to the spectrum. When the spectral lines and the teeth of the comb filter match the response is maximal. Nevertheless, experiments show that neither of these two models can fully explain human performance. Therefore, pitch perception models are still a field of active research.

The algorithm we presented in [1] combines information residing in the temporal and spectral representation and additionally suppresses spurious side peaks at harmonics and sub-harmonics as they occur in both rate and place models. Here we extend this algorithm by a pitch tracking based on grid based Bayesian tracking. We adapted the tracking algorithm from the formant tracking algorithm we proposed in [8]. Furthermore, in this paper we investigate how an additional multi channel preprocessing can be deployed to further increase the robustness of the algorithm. This preprocessing is based on a variant of blind source separation, namely Geometric Source Separation (GSS) [9]. Geometric Source Separation combines ideas from Blind Source Separation

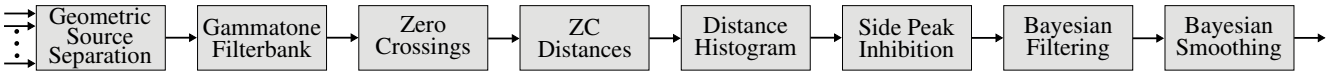

Fig. 1. System overview

and Beamforming to resolve common problems in Blind Source Separation as scaling and permutation by introducing geometric constraints on the microphone and sound source locations [10].

In the system we propose first GSS is used to separate the target signal from the interfering signals as background noise and noise produced by the robot (compare Fig. 1). Next the signal is transformed into the frequency domain with a Gammatone filter bank. In the following steps the zero crossings and based hereupon the zero crossing distances for each band pass signal are calculated. A histogram of these zero crossing distances is formed whereby side peaks resulting from harmonics and sub-harmonics of the true fundamental frequency are inhibited. Finally a grid based Bayesian tracking consisting of the steps of Bayesian filtering and subsequent Bayesian smoothing is applied.

In the following we will detail the building blocks of the proposed system for pitch extraction. After this we will give an overview on the human robot interaction scenario in which we tested our algorithm and evaluate the performance of the system based on a comparison of the tracking performance obtained on a clean headset signal and the signals recorded on the robot. We will conclude with a discussion of the results and an outlook on future work.

II. Geometric Source Separation

We used an online version of Geometric Source Separation (GSS) [9] for sound source separation. Since it allows easy addition and removal of sources with a small calculation cost, it is promising for mobile robots.

A spectrum vector of *M* sources and a spectrum vector of signals captured by the *N* microphones at frequency ω are denoted as $\mathbf{s}(\omega)$ and $\mathbf{r}(\omega)$, respectively. The spectrum vectors are obtained by application of the Fast Fourier Transform (FFT) on the time domain signals $\mathbf{s}(t)$ and $\mathbf{r}(t)$. The source separation is then formulated as

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{r}(\omega), \tag{1}$$

where $\mathbf{W}(\omega)$ is called a *separation matrix*. The separation is defined as finding $\mathbf{W}(\omega)$ which satisfies the condition that output signal $\mathbf{y}(\omega)$ is the same as $\mathbf{s}(\omega)$. In order to estimate $\mathbf{W}(\omega)$, GSS introduces two cost functions, that is, separation sharpness (J_{SS}) and geometric constraints (J_{GC}) defined by

$$J_{SS}(\mathbf{W}) = \|E[\mathbf{y}\mathbf{y}^H - \mathrm{diag}[\mathbf{y}\mathbf{y}^H]]\|^2 \tag{2}$$
$$J_{GC}(\mathbf{W}) = \|\mathrm{diag}[\mathbf{W}\mathbf{D} - \mathbf{I}]\|^2 \tag{3}$$

where $\|\cdot\|^2$ indicates the Frobenius norm, $\mathrm{diag}[\cdot]$ is the diagonal operator, $E[\cdot]$ represents the expectation operator and H represents the conjugate transpose operator. $\mathbf{D}$ is a transfer function matrix based on a direct sound path between a sound source and each microphone. $\mathbf{W}$ at the current time step t, $\mathbf{W}_t$, is estimated in an updating manner to minimize these cost functions as follows:

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \mu_{SS}\mathbf{J'}_{SS}(\mathbf{W_t}) + \mu_{GC}\mathbf{J'}_{GC}(\mathbf{W_t}).$$
$$\mathbf{J'}_{SS}(\mathbf{W_t}) = 2\mathbf{E}_{SS}\mathbf{W_t}\mathbf{r}\mathbf{r}^H$$
$$\mathbf{J'}_{GC}(\mathbf{W_t}) = \mathbf{E}_{GC}\mathbf{D}^H.$$

where $\mathbf{J'}(\mathbf{W})$ is an update direction of $\mathbf{W}$ derived from its complex gradient [11]. μ_{SS} and μ_{GC} are step-size parameters.

For further processing the source from the frontal direction is chosen and transformed back into the time domain via application of the Inverse Fast Fourier Transform (IFFT).

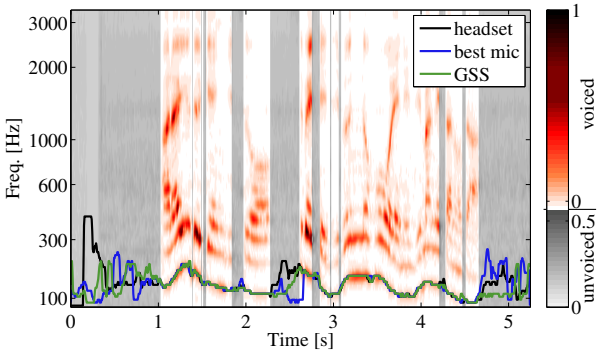

Fig. 2. Signal resulting from the GSS after application of the Gammatone filter bank. A male speaker is uttering the Japanese sentence "a r a y u r u g e N j i ts u w o s u b e t e j i b u N n o h o u e n e j i m a g e t a n o d a". The sentence contains a high proportion of vowels and voiced consonants and translates to "every fact was biased towards its preference". Pitch tracks for the clean and noisy signals are shown. Unvoiced regions are marked in gray.

III. Pitch Estimation

The algorithm we apply for pitch extraction relies on the calculation of a histogram of zero crossing distances and a subsequent inhibition of side peaks resulting form harmonics and sub-harmonics of the true fundamental frequency [1]. The first step of the pitch extraction is the transformation of the signal resulting from the GSS into the frequency domain via a Gammatone filter bank (see Fig. 2).

A. Extracting Rate Information

In most rate based systems the rate information is extracted via the autocorrelation function. However, the autocorrelation is very time consuming and not supported by biological data [12]. Therefore, we use in our system the zero crossing distances (ZCD) in the signal. Let $C_i = [t_{i,1}, t_{i,2}, \ldots, t_{i,N}]$ denote the ordered sequence of the time indices of all rising zero crossings, i. e. from negative to positive, in the band pass signals $g_i(n)$ in the i-th channel of the Gammatone filter bank:

$$C_i(m) = t_{i,m} \text{ with } g_i(t_{i,m}-1) < 0 \wedge g_i(t_{i,m}) \geq 0, \forall m. \tag{4}$$

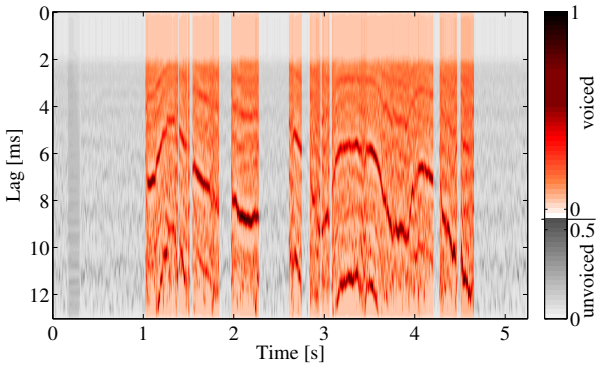

Fig. 3. Histogram of the zero crossing distances (ZCD) for a male utterance captured by the microphone with the highest SNR (best mic). Unvoiced regions are marked in gray.

Then the sequence of zero crossing distances is defined by

$$D_i(m) = C_i(m+1) - C_i(m). \tag{5}$$

Based on this a signal $d_i(t)$ is constructed which has in the interval between two zero crossings as its value the zero crossing distance. Hence $d_i(t) = D_i(m)$ where m is chosen such that $C_i(m) \leq t < C_i(m+1)$. This distance between adjacent zero crossings, more precisely its inverse, codes the frequency of the signal.

B. Zero Crossing Distance Histogram

When signals stem from the same fundamental frequency they have zero crossings in common. How many zero crossings they share depends directly on their harmonic order relative to the fundamental frequency. For example the first order harmonic shares each second zero crossing with the fundamental. Hence the distance between two zero crossings of the fundamental occurs again as the distance between three zero crossings of the first harmonic and so forth. We want to refer to these distances between multiple zero crossings as higher order zero crossing distances. Due to the frequency and articulation dependent phase delay introduced by the vocal tract not the absolute occurrence of the zero crossings is identical between harmonics of the same fundamental but rather their distance.

As a consequence of the reoccurrence of zero crossing distances of the fundamental in the harmonics a histogram of all distances shows a peak at the fundamental frequency. In Fig. 3 such a histogram is shown for the same utterance as depicted in Fig. 2. When interpreting the zero crossings as spikes of the neurons in the auditory system this histogram is very similar to a so called *all order interspike histogram* where a histogram of the phase locked firing of the neurons in the auditory system is calculated [13]. As not only the distances corresponding to the fundamental frequency but also those of the harmonics reoccur, the histogram shows many spurious side peaks corresponding to the harmonics and sub-harmonics of the true fundamental frequency. Sub-harmonics also occur because for instance the second order ZCD of the true fundamental frequency is also the first order distance of the first sub-harmonic ($\frac{1}{2}f_0$). This problem is not restricted to the histogram of the distances but does also occur when using the autocorrelation (compare [1]).

C. Extracting Place Information

The activity in the individual channels of the Gammatone filter bank codes the spectral information needed for place or pattern matching based pitch models. To implement a pattern matching algorithm we set up a comb filter for all possible fundamental frequencies with teeth at the harmonics $1 \ldots 15$. The range of possible fundamental frequencies is defined by the resolution of the zero crossing distances and hence by the sampling rate. At a sampling rate of $16\,\mathrm{kHz}$ a fundamental frequency of $80\,\mathrm{Hz}$ corresponds to 200 samples. The next possible fundamental frequency corresponds to 199 samples, $80.402\,\mathrm{Hz}$ respectively. In a scan through all possible fundamental frequencies from the lowest to the highest the corresponding comb filters are set up. For each of these comb filters the allocation of the teeth with harmonics of the current fundamental can be checked at each instant in time. The "filter response" of the comb filter is calculated based on the found allocation pattern. The better the found pattern matches the expected pattern the higher the response.

D. Combining rate and place information

One common way to determine the allocation of the teeth in the comb filter with harmonics is to use the energy in the band underlying the respective tooth. However, we propose to deploy the zero crossing distances previously calculated. The Gammatone filter bank has a limited frequency resolution due to a necessary trade off between filter bandwidth and settling time. A decrease in bandwidth and hence an increase in resolution comes at the cost of higher settling time which makes it impossible to analyze transient signals as speech. The ZCDs measure the instantaneous frequency in the time domain and hence are subject to this limitation to a lesser extent.

For each tooth of the comb filter the ZCD with the order corresponding to the harmonic order of the tooth is compared to the ZCD expected for the current fundamental frequency hypothesis. If the deviation between the expected and the measured distance is smaller than a predefined threshold t_Δ the tooth is said to be allocated by the expected harmonic. In the experiments reported later $t_\Delta = 4\%$. A modification of this method is not to check against the expected distance $\mathrm{ZCD}_{\mathrm{f}_0}$ but against the median of all distances found in the teeth. This yields slightly better results as the cross talk of the harmonics to adjacent filter bands delivers additional information.

E. Inhibition of Side Peaks

The creation of an allocation table for the comb filters allows to check the found allocation against expected ones. In Fig. 4 the allocation patterns are shown for the case where the hypothesis f_0' matches the true fundamental f_0, its first harmonic $f_0' = 2f_0$, or its first sub-harmonic $f_0' = \frac{1}{2}f_0$. These are the most important cases which produce spurious side peaks in the histogram. In order to distinguish the

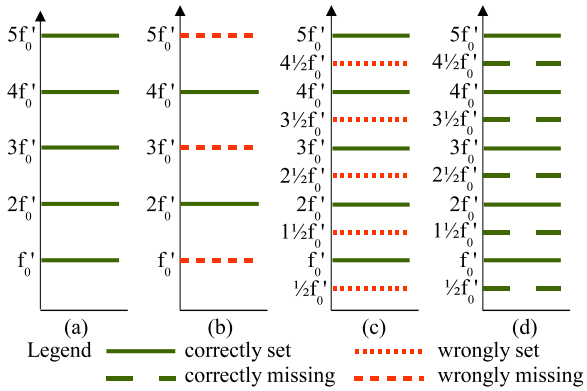

Fig. 4. Prototypical allocation patterns for the comb filters. In (a) the allocation is shown when the current hypothesis f_0' matches the true fundamental frequency f_0. In this case all teeth are set. The case where $f_0' = 2f_0$, hence the current hypothesis is in fact the first harmonic, is depicted in (b). Then only the teeth corresponding to the impair harmonic orders are set. When the current hypothesis $f_0' = \frac{1}{2}f_0$, i. e. the first sub-harmonic, all teeth are set but additionally also the teeth at $\frac{1}{2}f_0', (1+\frac{1}{2})f_0', (2+\frac{1}{2})f_0', \ldots$ are set. In order to capture this behavior the comb has to be extended by the sub-harmonics $(l+\frac{1}{2})f_0'$ (compare (c)). Plot (d) shows the case when the current hypothesis matches the true fundamental frequency for this extended comb filter.

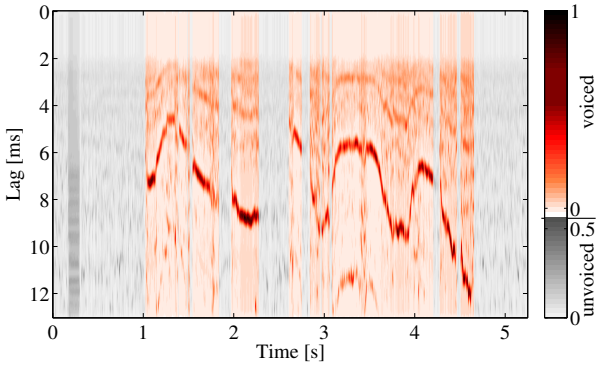

Fig. 5. Histogram of the zero crossing distances (ZCD) for a male utterance captured by the microphone with the highest SNR (best mic) after inhibition of side peaks.

case where the current hypothesis is the first sub-harmonic of the true fundamental frequency from that where the hypothesis is correct, additionally teeth at the sub-harmonics are included in the comb filter. By comparing the found allocation pattern to those producing most of the errors it is possible to inhibit the ones causing the errors. In the current implementation this is done by assigning a weight to each tooth such that the result is 1 if all and only the correct teeth are set and 0 if the found allocation pattern corresponds to the first harmonic or the first or second sub-harmonic ($f_0' = 2f_0 \vee \frac{1}{2}f_0 \vee \frac{1}{3}f_0$). The ZCD histogram is then only calculated for the fundamental frequency hypotheses. The weight with which they are entered in the histogram is determined by the response of the comb filter. This results in a canceling of the defined harmonics/sub-harmonics. In principle it is possible to extend this to more harmonics/sub-harmonics but the experimental results showed that this is not necessary. When comparing Figs. 3 and 5 one can see that this inhibition step makes the true fundamental frequency much better visible.

IV. Pitch Tracking

On the histogram of zero crossing distances we apply a tracking algorithm based on Bayesian filtering [8]. We originally developed this algorithm for formant tracking and adapt it in [14] to pitch tracking.

Bayesian filters represent the state at time t by random variables x_t, whereas uncertainty is introduced by a probabilistic distribution over x_t, called the belief $Bel(x_t) = p(x_t|z_1, \ldots, z_t)$. These filters sequentially estimate the beliefs over the state space conditioned on all information contained in the sensor data z_t [15].

Let $Bel^-(x_t)$ denote the predicted belief at time t which can be obtained via the application of the pitchs' underlying dynamics $p(x_t|x_{t-1})$. Then the belief at time t is calculated by correcting the predicted belief according to the observation from the pitch histogram $p(z_t|x_t)$ and a normalization factor α. Thus, the standard Bayesian filter recursion can be written as follows:

$$Bel^-(x_t) = \int p(x_t|x_{t-1}) \cdot Bel(x_{t-1})\, dx_{t-1} \quad (6)$$

$$Bel(x_t) = \alpha \cdot p(z_t|x_t) \cdot Bel^-(x_t) \quad (7)$$

Since we want to estimate pitch locations on a discrete grid defined by the evaluated zero crossing distance values, a grid-based approximation of the belief is chosen. Thus, assuming that N distances are evaluated, the state space at time t can be written as $X_t = \{x_{1,t}, x_{2,t}, \ldots, x_{N,t}\}$ which leads to the following Bayesian filter recursion:

$$Bel^-(x_{k,t}) = \sum_{l=1}^{N} p(x_{k,t}|x_{l,t-1}) Bel(x_{l,t-1}) \quad (8)$$

$$Bel(x_{k,t}) = \frac{p(z_t|x_{k,t}) Bel^-(x_{k,t})}{\sum_{l=1}^{N} p(z_t|x_{l,t}) Bel^-(x_{l,t})} \quad (9)$$

When operating in noisy conditions a subsequent backward pass on the already obtained filtering distributions $Bel(x_{k,t})$ is recommended since it significantly enhances the noise robustness of the algorithm. Bayesian smoothing provides such a mechanism. It aims to recursively estimate a smoothed version $\widehat{Bel}(x_{k,t})$ of the belief, thereby depending on both past and future observations [16]:

$$\widehat{Bel}(x_{k,t}) = p(x_{k,t}|z_1, z_2, \ldots, z_t, \ldots, z_{T-1}, z_T) \quad (10)$$

$$\widehat{Bel}^-(x_{k,t}) = \sum_{l=1}^{N} \widehat{Bel}(x_{l,t+1}) \cdot p(x_{l,t+1}|x_{k,t}) \quad (11)$$

$$\widehat{Bel}(x_{k,t}) = \frac{Bel(x_{k,t}) \cdot \widehat{Bel}^-(x_{k,t})}{\sum_{l=1}^{N} Bel(x_{l,t}) \cdot \widehat{Bel}^-(x_{l,t})} \quad (12)$$

The final calculation of exact pitch values $P(t)$ can easily be done by picking the peaks of the smoothed beliefs:

$$F(t) = \arg\max_{x_{k,t}} \left[\widehat{Bel}(x_{k,t})\right] \quad (13)$$

The result of the Bayesian tracking are overlaid on the spectrogram in Fig. 2 in blue (the two other curves result from the two other setups explained in detail in the following section).

We model the a priori distribution $p(x_{k,0})$ and the pitch

Fig. 6. 8 ch circular-type microphone array

dynamics $p(x_{k,t}|x_{l,t-1})$ with normal distributions.

V. Evaluation

The application scenario of our algorithm is the natural human robot interaction. This interaction will rely to a large extent on speech. One prerequisite for a natural interaction is in our opinion an interaction without the need for a headset. In our evaluation setting different people spoke to the Honda humanoid robot at a natural interaction distance of 1.5 m in a 4 m $\times$ 7 m room with RT_{20} = 300 ms[1]. 2 female and 6 male speakers were uttering a total of 90 utterances with 10-16 utterances per speaker. To enable a mutli-channel preprocessing the robot is equipped with an 8 channel microphone array (compare Fig. 6).

We compared the performance of our algorithm to a publicly available and commonly used pitch tracking frameworks get_f0 from ESPS in the implementation of the WaveSurfer toolkit [17]. This framework is based on an autocorrelation calculated from the full-band signal. It also includes a voicing detection and outputs pitch only for voiced segments. Because the voicing detection is rather unreliable for noisy speech we changed the parameterization such that the whole segment was classified as voiced and hence pitch was continuously calculated.

For the evaluation we also simultaneously recorded the speech signals with a headset. This headset signal was used to calculate the ground truth information for the fundamental frequency. The following results are given as deterioration of the tracking results relative to this assumed ground truth. The validity of the these results hence also partially depends on the correctness of the pitch extracted from the headset signal. However, visual inspection of the extracted pitch showed that the pitch is extracted very accurately from the headset signal. As pitch is only present in voiced regions of speech an additional voiced/unvoiced detection is necessary for the performance evaluation. To detect voiced regions we use the voicing detection algorithm described in [17]. The rationale behind this voicing detection is to decide that a segment is voiced if the normalized cross correlation $q_{\mathrm{NCCF}}(t)$, given by

$$q_{\mathrm{NCCF}}(t,\kappa) = \frac{1}{N} \frac{\sum_{j=t}^{t+N} r(j) r(j+\kappa)}{\sqrt{e(t)e(t+\kappa)}}, \qquad (14)$$

where $r(n)$ is the signal at time t and $e(t)$ its corresponding energy, is larger than a given threshold t_v:

$$q_{\mathrm{NCCF}}(t,\kappa) > T_v. \qquad (15)$$

In order to increase the robustness of the detection we additionally rejected segments with very low energy ($\approx 0.5\%$ of the mean energy). We applied this algorithm on the headset signal and used this information also for the noisy signals recorded on the robot. Consequently pitch tracking results were only evaluated in regions where voicing was detected in the headset signal.

After application of the GSS signals were downsampled to 16 kHz. In the implementation of the pitch tracking we used a 100 channel Gammatone filter bank with frequencies in the range from 80-5000 Hz . The implementation of the Gammatone filter bank is according to [18]. The range of possible fundamental frequencies was set to 80-500 Hz . We calculated zero crossing distances up to the order 7 and used a comb filter with 15 teeth. The Bayesian smoothing operated on a 100 ms time window.

To differentiate the impact of the multi-channel signal enhancement from the pitch extraction and tracking algorithm we compared two different setups. In the first setup we use the microphone signal with the highest SNR. As all speakers were speaking approximately from the front to the robot the SNR was always highest for the microphone mounted on the front (referred to in the following as *best mic*). A typical SNR value for this setup is $\approx$ 15 dB (compare to $\approx$ 35 dB for the headset)[2] In the second setup we evaluate the pitch tracking after the application of the GSS algorithm. The GSS improved the SNR $\approx$ 4 dB compared to the best mic condition.

In Table I the tracking errors relative to the headset signal are shown. The tracking performance of both algorithms in the noisy conditions is evaluated against the headset condition extracted by the same algorithm. Tracking errors are ceiled to 100%, i. e. errors larger than 100% are set to 100%.

TABLE I
Mean pitch tracking errors relative to headset signal in %.

	best mic	GSS	GSS+Post Filter
get_f0	2.6	7.1	7.2
proposed	2.1	1.5	2.2

Additionally, we also evaluated the so called Gross Pitch Error (GPE) [19]. It measures how much of the pitch track deviates more than e_t from the true pitch. In our case we set $e_t = 20\%$. The corresponding values are given in Table II.

The results show that the tracking errors already for the best mic configuration are very good for both algorithms. The GSS preprocessing notably reduces the errors for our algorithm. However, the results for get_f0 were deteriorated

[1] RT_{20} is better suited for measurements in noisy environments. It gives the decay measured at 20 dB extrapolated to 60 dB decay

[2] We calculated the SNR as the ratio of the energy of the segments containing only speech to those containing only noise. Signal distortions due to reverberations are hereby not taken into account.

TABLE II
GROSS PITCH ERRORS ($> 20\%$) RELATIVE TO HEADSET SIGNAL IN %.

	best mic	GSS	GSS+Post Filter
get_f0	1.8	2.7	2.9
proposed	0.7	0.3	1.0

by the GSS. When using the GSS as preprocessing combined with our algorithm the errors are very small and only very little gross pitch errors occur.

The GSS based signal enhancement proposed in [9] also includes a multi-channel post filtering step. The post filter is applied after the GSS and has as its purpose to reduce the noise still present after the GSS step. In addition to the stationary components of the noise it also estimates non-stationary components and subtracts them from the signal. We investigated a setup where we included the post filter as described in [9]. When comparing Table I and II one can see that the post filtering is not beneficial for the pitch tracking for all algorithms.

VI. SUMMARY & DISCUSSION

We developed a system which is able to extract the fundamental frequency of a speaker in natural human robot interaction, i. e. without the use of a headset. The main building blocks of the system are a multi-channel preprocessing based on Geometric Source Separation (GSS), pitch extraction based on a zero crossing distance histogram, and pitch tracking using a grid based Bayesian tracker.

We evaluated the system for 8 different speakers each uttering 10-16 sentences. The results showed that the pitch extraction and tracking already yields good results without the preprocessing but that the results could further be improved with the GSS. The application of an additional post filter decreased performance. We attribute this to the fact that on one hand the pitch extraction does not depend on the amplitude of the signals and hence the possibly beneficial effects of the post filtering can not be exploited by the pitch extraction. On the other hand the distortions following from the post filtering, e. g. musical tones resulting from incorrect estimation of either noise or signal energy, impair the pitch extraction.

The comparison of our algorithm to get_f0 from the Snack toolkit showed that our algorithm performs significantly. In the best case, i. e. using our algorithm followed by the GSS but without post filtering, we obtain relative errors averaged over all speakers below 2% and gross pitch errors of only 0.3%. From this we conclude that the system we propose robustly extracts the fundamental frequency and hence lays the foundation for a prosodic analysis of the speech signal.

The results show that the GSS is beneficial for the pitch extraction in Human-Robot interaction. Up to now we only performed both algorithms in a sequence. In the future we will investigate how a tighter integration can be obtained, e. g. by replacing the FFT and IFFT step in the GSS by a Gammatone filter bank which will avoid repeated transformations from the time domain in the frequency domain as in the current system.

VII. ACKNOWLEDGMENTS

We want to thank Dr. Shun'ichi Yamamoto for support with the GSS algorithm and for designing and performing the recordings and Claudius Gläser for providing the Bayesian tracking algorithm.

REFERENCES

[1] M. Heckmann, F. Joublin, and C. Goerick, "Combining rate and place information for robust pitch extraction," in *Proc. INTERSPEECH*, Antwerp, 2007, pp. 2765–2768.

[2] A. de Cheveigne, "Pitch perception models," in *Pitch*, C. Plack and A. Oxenham, Eds. Springer, Cambridge, U.K., 2004.

[3] J. R. C. Licklider, "A duplex theory of pitch perception," *Eperientia*, vol. 7, pp. 128–134, 1951.

[4] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. on Acoustics Speech and Signal Proc.*, vol. 24, no. 3, pp. 201 – 212, 1976.

[5] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I pitch identification," *Journal of the Acoust. Soc. Am.*, vol. 89, pp. 2866–2882, 1991.

[6] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C.Zhang, and M. H. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception, Proceedings of the 9h International Symposium on Hearing*, Y Cazals, L. Demany, and K. Horner, Eds., Pergamon, Oxford, 1992, pp. 429–446.

[7] J. L. Goldstein, "An optimum processor theory for the central formation of the pitch of complex tones," *Journal of the Acoust. Soc. of America*, vol. 54, pp. 1496–1516, 1973.

[8] C. Gläser, M. Heckmann, F. Joublin, and C. Goerick, "Combining auditory preprocessing and bayesian estimation for robust formant tracking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 224–236, 2010.

[9] S. Yamamoto, K. Nakadai, J.M. Valin, J. Rouat, F. Michaud, K. Komatani, T. Ogata, and HG Okuno, "Making a robot recognize three simultaneous sentences in real-time," in *Proc IEEE/RSJ Int. Conf. on Robots and Intell. Syst. (IROS)*, Edmonton, Canada, 2005, pp. 4040–4045.

[10] L.C. Parra and C.V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. on Speech and Audio Proc.*, vol. 10, no. 6, pp. 352–362, 2002.

[11] D.H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proc.*, vol. 130, no. 1, pp. 251–276, 1983.

[12] C. Kaernbach and L. Demany, "Psychophysical evidence against the autocorrelation theory of auditory temporal processing," *Journal of the Acoustic. Soc. of America*, vol. 104, pp. 2298–2306, 1998.

[13] P. A. Cariani, "Temporal codes and computations for sensory representation and scene analysis," *IEEE Trans. Neural Networks*, vol. 15, pp. 1100–1111, 2004.

[14] M. Heckmann, C. Gläser, M. Vaz, T. Rodemann, F. Joublin, and C. Goerick, "Listen to the parrot: Demonstrating the quality of online pitch and formant extraction via feature-based resynthesis," in *Proc. IEEE/RSJ Int. Conf. on Intell. Robots and Systems (IROS)*, Nice, 2008, IEEE-RSJ.

[15] D. Fox, J. Hightower, L. Liao, D. Schulz, and G. Borriello, "Bayesian Filtering for Location Estimation," *IEEE Pervasive Computing*, vol. 2, no. 3, pp. 24–33, 2003.

[16] S. J. Godsill, A. Doucet, and M. West, "Monte Carlo smoothing for nonlinear time series," *J. of the American Stat. Assoc.*, vol. 99, no. 465, pp. 156–168, 2004.

[17] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

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

[19] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 24, no. 5, pp. 399–418, 1976.