Sound Source Separation for a Robot Based on Pitch

Martin Heckmann, Frank Joublin, Edgar Körner

2005

Preprint:

This is an accepted article published in Proceedings of the International Conference on Itelligent Robots & Systems (IROS). The final authenticated version is available online at: https://doi.org/[DOI not available]

Sound Source Separation for a Robot Based on Pitch

Martin Heckmann, Frank Joublin, Edgar Körner
Honda Research Institute Europe GmbH
Carl-Legien-Strasse 30
D-63073 Offenbach/Main, Germany
{martin.heckmann, frank.joublin, edgar.koerner}@honda-ri.de

***Abstract*— We present a novel method for the separation of monaurally recorded speech signals based on pitch. Our method is inspired by the ability of some auditory neurons to phase lock with the excitation signal. After applying a Gammatone filterbank on the original signal we compare the distances between zero crossings of possible harmonics and decide upon the result of this comparison if they share the same fundamental and hence originate from the same sound source. For higher frequencies we use the amplitude modulation property of unresolved harmonics to determine their fundamental frequency. When comparing our method to standard autocorrelation based methods we see that the pitch can be tracked more precisely and especially opens the way to extract also the pitch contour of a second speaker or other sound sources which can be of importance for the robots behavior. Tests in sound source separation of our algorithm on a database with several speakers and a large set of intrusions show that our algorithm performs slightly better than the commonly used autocorrelation at lower computational costs.**

***Index Terms*— Monaural Sound Source Separation, Zero Crossing Distances, Histogram, Amplitude Modulation**

I. Introduction

The interaction with a robot like Asimo by speech is especially difficult due to the long distances between the speaker and the microphones installed on the robot. Therefore, additionally to the desired speech signal interfering speech signals and background noise are captured. Hence sound source separation is an important aspect in robot audition. Todays speech recognition systems show dramatic performance impairments in such scenarios. However, it is well known that humans can cope with such situations surprisingly well [1]. This finding has stimulated tremendous research in the domain referred to as *Computational Auditory Scene Analysis (CASA)*, which sets itself as a target to achieve human performance by computational means [2], [3], [4].

In particular the goal of our algorithm is to separate speech signals in monaural recordings even in very adverse conditions when significant background noise and additional speakers are present at the same time. Particularly we try to decide for each time frequency region which of the different sound sources dominates and then build for each sound source a binary mask which is one at those time frequency regions where the sound source dominates and zero at the others. This makes the assumption that only those regions should be retained where the desired signal dominates and all others should be rejected, an assumption commonly made in CASA like systems and leading to very good separation results [5], [6]. The separation in our algorithm is based on common fundamental frequency, whose percept is called pitch [7]. A separation based on fundamental frequency is only possible in voiced speech segments. To demonstrate the performance of our algorithm we therefore use completely voiced sentences.

II. Model Overview

The first step in our sound source separation system is the division of the input signal into different frequency bands via a Gammatone filterbank.The aim of the separation is to label the different frequency channels for each instant in time with the pitch they emanate from. Our algorithm does not impose blocks in the time domain but they rather arise signal driven. Nevertheless we want to use the term *time-frequency (TF)* units for these blocks in a given channel. Once all channels are labeled the possible sound sources in the signal have to be identified and their evolution over time has to be tracked. In the case of voiced segments the different sound sources can then be separated by grouping channels with a common pitch.

For the implementation we used a 128 channel Gammatone filterbank with frequencies in the range from 80-5000 Hz. The implementation of the Gammatone filterbank is according to [8]. In order to reduce the phase distortions introduced by the filterbank in each channel the delay for the center frequency is calculated and a delay is introduced so that at the end all center frequencies have the same delay. The range of possible fundamental frequencies for our algorithm was set to 80-500 Hz, which is in accordance with the pitch variation found in the database used for testing. However this is not a limitation of the proposed algorithm an can easily be extended to allow e.g. also music perception.

III. Pitch Estimation

The starting point for the sound source separation is the determination of the fundamental frequencies of the signals present in the acoustic scene. Based on these fundamental frequencies the TF units can then be labeled and allocated to the corresponding sound source. Most pitch estimation algorithms in the literature are based on the autocorrelation function. Either on autocorrelation of the complete signal or

by combining the autocorrelations of different channels [9]. It is noteworthy that the subdivision in different channels only leads to an improvement if a nonlinear processing is performed for each channel [9].

The autocorrelation is very time consuming and biologically rather implausible. Therefore, we propose to use a different mechanism to determine if two filter output signals originate from one common fundamental frequency. Our approach is inspired by the phase locking property of some neurons in the auditory system and relies on the distances of the zero crossings of the filter signals. These distances are evaluated for each channel of the Gammatone filterbank and then compared over different channels.

A. Using zero crossings

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 crossings. 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 (compare Fig. 1). We use only the

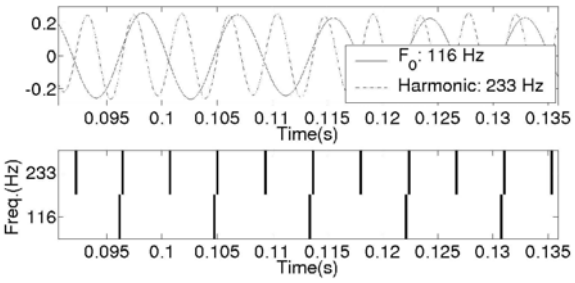

Fig. 1. Visualization of the identical distance between zero crossings for the fundamental and its first harmonic. In the upper plot the fundamental frequency in the 116 Hz channel and its first harmonic in the 266 Hz channel are shown. The lower figure only shows their zero crossings visualized by bars. In the lower half of the plot the zero crossings for the 116 Hz channel and in the upper half those of the 266 Hz channel are displayed.

zero crossings with positive slope (from negative to positive) but those with negative slope could be used as well. The reciprocal of the zero crossing distance in time is of course the frequency of the signal under investigation. In this sense this distance can be used to measure the frequency of the signal. The measurement is performed completely in the time domain but can be only updated with each period of the signal. Therefore a signal driven blocking of the signals determined by the zero crossing distances arises.

From biological studies it is known that certain neurons in the auditory system fire in phase with the basilar membrane movement, which is denominated as *phase locking*. The zero crossings are a way to mimic this phase locked spike firing.

B. Zero crossing distance histogram

As the distance of the fundamental reoccurs in the higher order distances of the harmonics a histogram over all distances shows peaks at the distance value of the fundamental. The energy of the filter signals is represented in the histogram by weighting the values with the energy. This means distance values stemming from a TF unit with high energy have more weight in the histogram. In order to further enhance the formation of peaks at the fundamental, different weights are put on the different orders of the zero crossings. The weighting function chosen is $1/(n+1)$, where n is the order with the fundamental having the order 0. This takes into account that usually the lower order harmonics have more energy than the higher ones. Additionally a comb filter is used in the calculation of the histogram. In a loop over the range of possible zero crossing distances of the fundamental only the channels whose center frequencies can be in a harmonic relation to the fundamental currently under investigation are used for the calculation of the histogram. To allow for some overlap in the channels each comb in the comb filter has a width of 3 channels. In Fig. 3 a) the resulting histogram for the mixture of two male speakers, where one utterance is completely voiced, displayed in Fig. 2, is shown. The maxima in the distance direction clearly

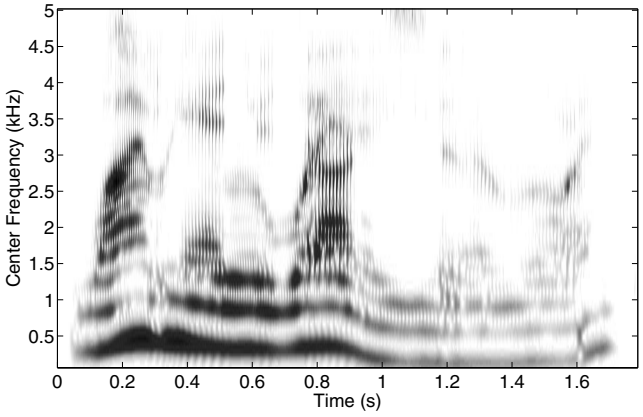

Fig. 2. Amplitude of the envelope at the output of the Gammatone filterbank of the mixture of an all voiced male foreground utterance and a second male utterance as intrusion. The SNR of the signal was 7dB.

correspond to the zero crossing distances of the fundamental of the foreground utterance. Additionally side maxima occur at harmonics of the fundamental. These side maxima can be eliminated when the main maximum is determined and hence

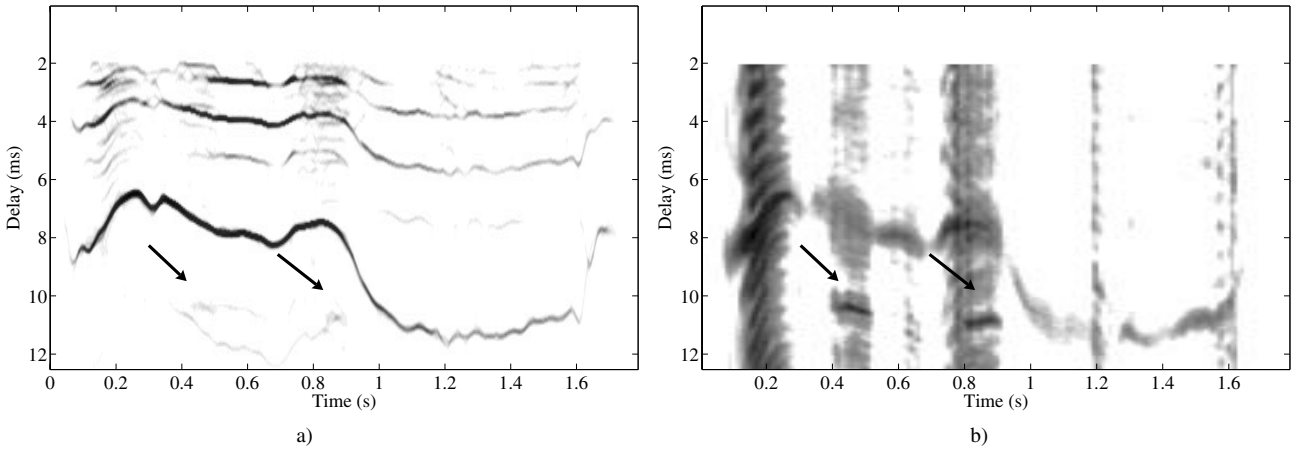

Fig. 3. Zero crossing distance histogram of the signal displayed in Fig. 2 is shown in a). The peak found at the fundamental is clearly visible. Additional peaks from the voiced parts of the background utterance, indicated by arrows, can be seen. The harmonics of the fundamental form parallel curves to the correct fundamental. In b) the sum of the autocorrelation function over all channels for the same sound file generated by the implementation in [6] which additionally implements a model of the outer hair cells is shown. The second sound source is completely covered by side maxima.

the saliency of the second utterance can be enhanced. When comparing the distance histogram to the widely used sum of the autocorrelation function (e. g. [6]) in Fig. 3 b) it can be seen that the fundamental of the foreground utterance is much less visible as in the distance histogram and the fundamental of the second sound source completely vanishes.

C. Unresolved harmonics

For high frequencies the zero crossings get very close together and hence can only be used in a limited way to identify the pitch. This problem can be solved by exploiting a property of the filters of the Gammatone filterbank [6]. The filters have, similar to the filters formed by the basilar membrane in the human ear, a constant relative bandwidth. As a consequence for higher harmonics several harmonics fall into one filter channel, which is termed as unresolved harmonics in opposition to low order harmonics, where only one harmonic is present in a filter channel [10], [7]. Due to the coherent interaction of these unresolved harmonics the resulting signal shows amplitude modulation with the underlying fundamental. Therefore we demodulate the signal by rectification and low pass filtering. The cut off frequencies of the low pass filters are set to the minimum between half the center frequency of the underlying Gammatone filter and twice the maximum pitch value. In order to determine the modulation frequency we feed the envelope signal in a second Gammatone filterbank with identical bandwidth as the first one but where only the channels up to twice the maximum pitch are used (compare Fig. 4). For each of these signals the first and second order zero crossing distances are calculated. Based on these a histogram as described in Sec. III-B is calculated, where only the first and second order zero crossing distance is used. Here the property that the energy of the modulation envelope is mainly concentrated

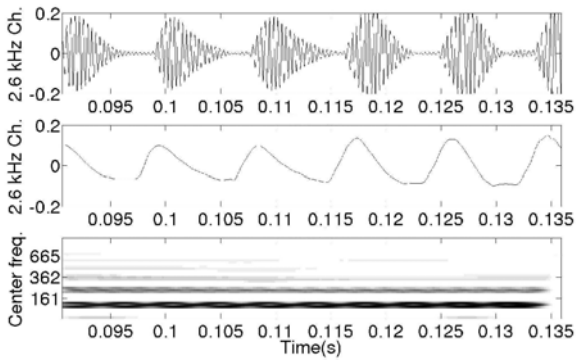

Fig. 4. The middle plane shows the result of the demodulation of the unresolved harmonic in the top plane. In the bottom plane the effect of the application of the second filterbank on the envelope signal in the middle plane is shown. The main peak at the modulation frequency as well as the harmonics of it are clearly visible in the bottom plane.

on the fundamental frequency and to a lesser degree on its first harmonic is used. From the maximum in the histogram the fundamental of the underlying unresolved harmonic can be determined. We want to refer to this algorithm to extract the zero crossing distances as *Amplitude Modulation (AM) criterion*. The resulting distance value weighted with the energy of the channel is added to the distance histogram described in Sec. III-B. Additionally values obtained this way from unresolved harmonics are weighted with 1/3 taking into account that unresolved harmonics usually carry less information about pitch than low order resolved harmonics [7]. The distance histogram in Fig. 3 is the result of the combination of the distance values from resolved and unresolved harmonics.

When white noise at very low SNR levels is present as intrusion the distances calculated from the amplitude modulation show a high variation and reflect in most parts only the

distances of the noise. To overcome this problem we calculate the variance of the distance values and reject those with a high variance as in speech segments the distances change only rather slowly. The variance is calculated by a high pass filtering of the distance values. The cut-off frequency of the high pass was set to 20 Hz. The resulting signal is demodulated by rectification and low pass filtering where the low pass also has a cut-off frequency of 20 Hz. When this distance envelope signal has an amplitude of more than 0.7 ms the corresponding sample is rejected. This procedure largely rejects distance values stemming from white noise and leaves the remaining values mostly unchanged. Consequently the tracking of the pitch in white noise is improved and works even at SNR levels of -8 dB successfully.

In contrast to [6] we also use the unresolved harmonics for the determination of pitch. Therefore our algorithm is also able to detect pitch correctly for speech parts where all harmonics are unresolved as it is the case when people speak with very low pitch.

D. Pitch tracking

From the beforehand calculated zero crossing distance histogram now the course of the pitch over time can be tracked. A critical decision in the tracking is which sound source should be tracked. As the focus of this article is not the actual tracking of the pitch but rather to develop a new method to determine the pitch and segregate sound sources based on pitch, we make some restricting assumptions during the tracking. The strongest assumption is that the desired sound source is all voiced. Therefore we know that the desired pitch track has no interruptions. Firstly we calculate the maxima in the distance histogram and build the five longest segments of connected maxima, where connected means that the distance value does not change more than 5% from one sample to the other. As the desired pitch does not always correspond to the maximum in the distance histogram these previously determined tracks only span part of the utterance. Next we grow each of these five tracks at either end. For this purpose we weight the values in the histogram with a Gaussian like window centered at the distance position corresponding to the linear extrapolation of the slope of the track. By doing so we enhance distance values in the expected direction of the track and reduce those outside this direction. This slope is calculated from the previous 20 ms of the track. The actual window is given by:

$$w = 1 + 6 \cdot \exp\left(-\frac{\hat{d}^2}{2\sigma^2}\right). \tag{1}$$

The Gaussian in Eq. 1 is centered at the distance value in the expected direction $\hat{d}$ where σ is calculated according to

$$\sigma = \sqrt{\frac{0.1\hat{d}}{25}}, \tag{2}$$

All values outside the 5% search region region around $\hat{d}$ are set to zero. The growing of the tracks finishes when no values in the search region are above a threshold defined by twice the mean of the histogram. Tracks spanning less than 60% of the utterance are rejected. For each remaining track the sum of all histogram values covered by the track is calculated and then divided by the length of the track. The final pitch track is the one where this mean histogram value is the highest.

IV. Pitch Labeling

A crucial step in the desired sound source separation is the labeling of the TF units with the fundamental frequency they originate from so as to later group TF units according to common fundamental frequency. The common approach for the labeling of TF units with pitch is again the use of the autocorrelation of the TF unit under concern [6].

As for the determination of the pitch, we also use for the labeling of the TF units the zero crossing distances. The closer the zero crossing distance of the channel under investigation is to that found for the found pitch, the more likely it is that it belongs to the same sound source. This difference can hence be mapped to a reliability measure. We calculate these difference values independently for distances obtained directly from the filter signal and those obtained by the amplitude modulation criterion. As both criteria are evaluated over the full frequency range, we do not classify TF units into resolved or unresolved harmonics a priori but rather make this decision based on which of the two criteria shows the minimum difference to the distance value of the found pitch. In the case of the resolved harmonics the resulting difference value is the minimum of the difference between the zero crossing distance of the found pitch and the distances calculated at all orders of zero crossing distances. For the unresolved harmonics the resulting difference value is simply calculated as the difference between the zero crossing distance of the found pitch and the distance calculated from the AM criterion. When introducing a threshold we can make a decision if a given TF unit belongs to the found pitch or not. This threshold can be set independently for the resolved and unresolved harmonics.

In Fig. 5 a) the mask resulting from the labeling is shown. The mixture is the same as in Fig2 containing a male target utterance and a male utterance as intrusion. Black regions were grouped to target speech and white regions to the intrusion. In Fig. 5 b) also the ideal binary mask is shown. Given that we have the signals before mixture at our disposal we can calculate for each instant in time in the mixture if the target speech or the intrusion has more energy. Regions where the target speech dominates are marked in black and those where the intrusion dominates in white. By comparing the two masks it can be seen that our algorithm correctly assigns most parts of the signal to either target or intrusion.

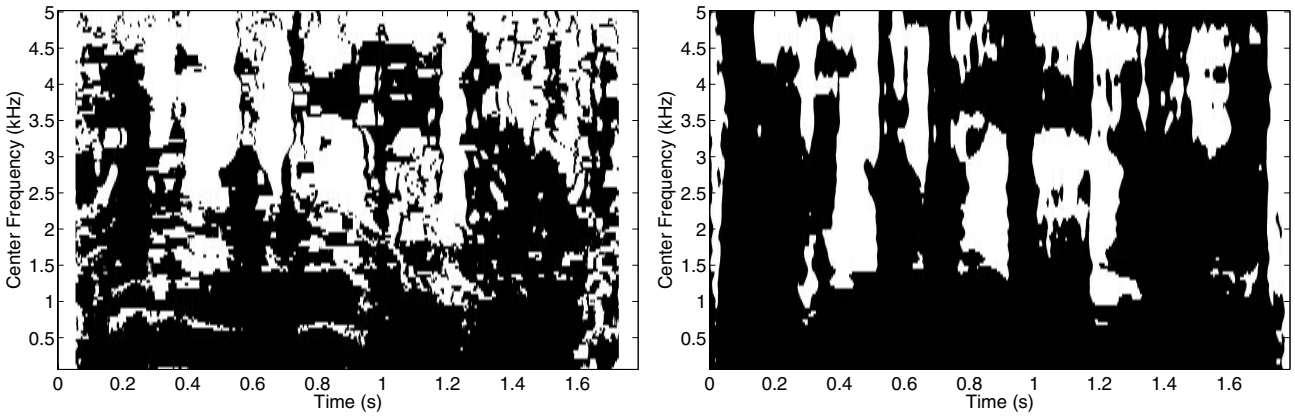

Fig. 5. In a) the mask generated by our algorithm is shown, where black regions are assigned to target speech and white regions to the intrusion. The ideal binary mask is visualized in b), where dominance of target speech is represented by black and white regions indicate the dominance of the intrusion.

V. Results

In order to asses the performance of our algorithm we evaluated it on a database collected by Cooke which was frequently used to test CASA systems [3], [11], [6]. The database consists of 10 utterances of a male speaker mixed with 10 intrusions yielding a total of 100 utterances [2]. All of these utterances are completely voiced (e. g. 'Why were you all weary') and hence have a continuous pitch. To these utterances different intrusions were mixed. Amongst these

TABLE I
List of the signals used as intrusions in the Cooke database

Name	Type	Name	Type
N0	1 kHz pure tone	N5	siren
N1	white noise	N6	trill telephone
N2	noise bursts	N7	female speech
N3	'cocktail party' noise	N8	male speech
N4	rock music	N9	female speech

are white noise, music, and male and female speakers (see Tab. I for a full list). The sampling rate of the database is 16 kHz. For resynthesis of the signals we first undo the phase compensation of the center frequencies and then invert the filter signals in time, pass them again through the Gammatone filterbank and then invert them again in time. By this backward filtering the phase delay introduced by the Gammatone filterbank is compensated [12].

The maximum order of zero crossings of resolved harmonics was set to 7. Distance thresholds for the acceptance or rejection of a TF unit were set to 0.3125 ms (or 5 samples) for resolved and 0.625 ms (or 10 samples) for unresolved harmonics.

As performance measure we have chosen the *Signal to Noise Ratio (SNR)*. Enven though the SNR has its drawbacks when measuring CASA system performance it is still commonly used. We measured the SNR of the output signal as follows:

$$\mathrm{SNR}_{\mathrm{res}}(\mathrm{V},\mathrm{N}) = 10\log\left(\frac{\sum_{k=0}^{K} v_{\mathrm{V}}(k)^2}{\sum_{k=0}^{K}\left(v_{\mathrm{V}}(k) - r_{\mathrm{V,N}}(k)\right)^2}\right), \quad (3)$$

where v_{V} is the original target utterance and $r_{\mathrm{V,N}}(k)$ the target utterance reconstructed from the mixture of utterance V with intrusion N. The SNR enhancement is defined as the difference between the resulting SNR and that of the original mixture. The SNR values of the original mixture, the resulting signal after separation and those of the system by Hu and Wang [6] averaged over all speakers for a given intrusion are given in Tab. II. In order to make results more comparable in the case of the Hu and Wang system also no segmentation was used. As can be seen the two models perform in average similar, but the proposed algorithm achieves overall an improvement of 9%. Due to the high standard deviations this performance difference is not statistically significant, though the standard deviation also is 9% lower for our algorithm (compare Tab. II). For some intrusions the differences are rather big, but no clear trend can be identified where one or the other algorithm works better as no systematic relation between the intrusions where this happens could be identified.

VI. Discussion

We presented a novel method to identify the pitch in a mixture of sound sources and to separate the sound sources based thereupon. The presented algorithm is inspired by biological and psychological mechanisms. An important difference to the mainly used autocorrelation function is that we obtain a constant resolution over frequency. Independent of signal frequency this resolution is only determined by the distance between the sampling points. In contrast to this the resolution of the autocorrelation is only weak for

TABLE II

SNR RESULTS IN DB AND STANDARD DEVIATION (SD) FOR AVERAGE

Intrusion	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	Average	(SD)
Mixture	-4.3	-1.9	12.5	4.2	3.9	-6.7	3.5	7.1	12.9	5.9	3.7	(6.5)
Hu-Wang system	10.0	4.3	15.1	8.0	8.6	4.1	10.0	11.5	16.4	7.4	9.5	(4.6)
Our Model	8.2	6.8	15.4	8.3	10.4	3.8	15.5	11.3	14.7	9.8	10.4	(4.2)

low frequencies as low frequencies form wide peaks in their autocorrelation function. Additionally, our algorithm does not work with predefined blocks, rather blocks emanate signal driven with the fundamental period of the signal under investigation. This way changes in the fundamental frequency can be much easier tracked as they also directly influence the analyzing block length. Also the calculation of the zero crossing distances is less costly as the autocorrelation. We can not provide any clear numbers here as our algorithm is only partly implemented in C, but it seems that an improvement of factor 3-5 in calculation time is likely. Furthermore the autocorrelation is physiologically rather implausible whereas a simple zero crossing detection and integration seems much more likely.

The zero crossing distance is a robust measure as we apply it to bandpass signals, resulting from rather narrow bandpass filters. This way noise overlaid to the signal is filtered out and the resulting signal is very close to pure sinusoidal signals. A detailed analysis of the noise robustness of the zero crossings is given in [13].

The aforementioned noise reduction property of the bandpass filterbank is what we also utilize for the treatment of unresolved harmonics. By applying an identical filterbank as used for the original signal on the modulation envelope of unresolved harmonics we spread the noise over different channels whereas the modulation bearing channel retains the largest energy and is clearly identifiable.

The introduced zero crossing distance histogram takes all harmonics up to a certain order of a signal into account. This way, similar to humans, the fundamental can also be tracked, when the actual fundamental is not present [7]. When comparing the distance histogram to the summed autocorrelation function the actual pitch is much more visible in the distance histogram. This is why a rather simple tracking algorithm in contrast to the multi stage tracking algorithm used in [6] suffices to track the target signal. Especially additional sound sources are much better observable in the distance histogram as in the autocorrelation.

The very challenging problem of tracking the pitch contour was not in the focus of this paper though. We made, in a similar way to [6], simplifications for the tracking and adapted it to the database used for testing. In a realistic scenario voiced segments alternate with unvoiced segments, where no pitch is present, making the tracking quite difficult. To solve this problem additional cues like sound source localization and onsets are certainly necessary.

We further showed that the zero crossing distances can also be used efficiently to separate the sound sources once the pitch is identified. For this purpose we used a database with different speakers with a wide variety of intrusions. The separation results by our algorithm were slightly, but not significantly better than those obtained by the autocorrelation.

Recapitulating we showed that zero crossings, being biologically more plausible, can be used very efficiently to identify pitch and separate sound sources based on pitch. The resulting algorithms are much simpler as those commonly deployed when using the autocorrelation function and lead in the case of the estimation of the pitch of multiple sound sources to significantly better results.

REFERENCES

[1] A. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

[2] Martin Cooke, *Modeling Auditory Processing and Organization*, Ph.D. thesis, Cambridge University, Cambridge, U.K., 1993.

[3] D. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, MIT Department of Electrical Engineering and Computer Science, 1996.

[4] Martin Cooke and Daniel P.W. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 33, pp. 141–177, 2000.

[5] Martin P. Cooke and Phil Green, "Robust automatic speech recognition with missing and unreliable data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[6] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. On Neural Networks*, vol. 15, pp. 1135–1150, 2004.

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

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

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

[10] H. Helmholtz, *Die Lehre von den Tonempfindungen*, Vieweg, Braunschweig, 1863.

[11] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. On Neural Networks*, vol. 10, pp. 684–697, 1999.

[12] M. Weintraub, *A theory and computational model of auditory monaural sound separation*, Ph.D. thesis, Stanford Univ., 1985.

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