

Monaural Speech Segregation Based on Pitch

Martin Heckmann, Frank Joublin, Christian Goerick

Honda Research Institute Europe GmbH, 63073 Offenbach am Main, Deutschland,

Email: {martin.heckmann, frank.joublin, christian.goerick}@honda-ri.de

Introduction

The goal of the proposed 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. The separation in our algorithm is based on common fundamental frequency, whose percept is called pitch [1]. 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. The first step in our sound source separation system is the division of the input signal into different frequency bands via a Gammatone filterbank. 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 [2]. All filters are set to have equal phase delay. The range of possible fundamental frequencies for our algorithm was set corresponding to the database used for testing to 80-500 Hz.

Pitch Estimation

Most pitch estimation algorithms in the literature are based on the autocorrelation function [4]. 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.

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. Hence the distance between two zero crossings of the fundamental occurs again as the distance between four 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. We use only the zero crossings with positive slope. 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 firing.

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. 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. Additionally a comb filter is used in the calculation of the histogram. In Fig. 1 a) such a histogram is shown. The maxima in the distance direction clearly correspond to the zero crossing distances of the fundamental of the foreground utterance. Additionally side maxima occur at harmonics of the fundamental, which can be eliminated once the fundamental is determined. The comparison of the distance histogram to the widely used sum of the autocorrelation function (e.g. [3]) in Fig. 1 b) shows that the fundamental of the foreground utterance is much less visible in the autocorrelation as in the distance histogram and the fundamental of the second sound source completely vanishes, showing that zero crossings are better suited to extract pitch for multiple sources.

Unresolved harmonics: When multiple harmonics fall into one filter channel, termed unresolved harmonics, their coherent interaction leads to amplitude modulation of the resulting signal with the underlying fundamental frequency. In order to extract this modulation we demodulate the signal by rectification and low pass filtering and then feed the envelope signal in a second Gammatone filterbank. A histogram as described before is calculated with only the first and second order zero crossings. From the maximum in the histogram the fundamental of the underlying unresolved harmonic can be determined. The resulting distance value weighted with the energy of the channel is added to the distance histogram. Distance values with high variation are rejected. This improves the tracking of the pitch in white noise so that it even works at SNR levels of -8 dB successfully.

Pitch tracking: From the beforehand calculated zero crossing distance histogram now the course of the pitch over time can 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, hence pitch is uninterrupted. 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. Next we grow each of these

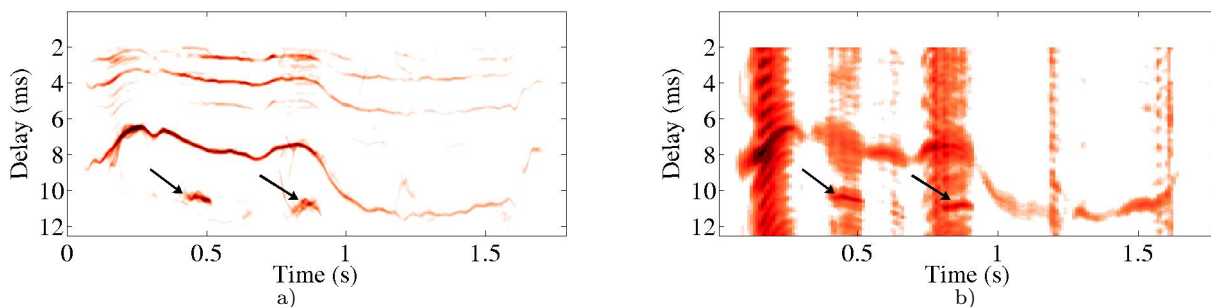


Figure 1: Zero crossing distance histogram of two male speakers is shown in a). The peak found at the fundamental is clearly visible. Additionally peaks from the voiced parts of the background utterance, indicated by arrows, can be seen. In b) the sum of the autocorrelation function over all channels for the same sound file generated by the implementation in [3] which additionally implements a model of the outer hair cells is shown. The second sound source is completely covered by side maxima.

Table 1: SNR results in dB.

Intrusion	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9	Average
Mixture	-4.32	-5.17	8.72	3.90	3.78	-6.90	1.48	6.94	12.63	5.84	2.69
Our Model	8.20	6.81	15.35	8.26	10.36	3.76	15.50	11.28	14.66	9.77	10.40
Hu-Wang system	10.01	4.29	15.12	7.95	8.62	4.11	10.05	11.48	16.36	7.38	9.54

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. All values outside the 5% search region 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. The final pitch track is the one which covers the highest energy in the histogram relative to its length.

Pitch Labeling

The closer the zero crossing distance of the channel under investigation is to that of 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 from the amplitude modulation. 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 amplitude modulation. When introducing a threshold we can make a decision if a given segment belongs to the found pitch or not.

Results

We evaluated our algorithm on a database collected by Cooke consisting of 10 utterances of a male speaker mixed with 10 intrusions, a 1 kHz pure tone (N0), white noise (N1), noise bursts (N2), 'cocktail party' (N3), rock music (N4), a siren (N5), a trill telephone (N6), female speech (N7, N9), and male speech (N8), yielding a total of 100 utterances [5]. All utterances are completely voiced. The

sampling rate of the database is 16 kHz. The maximum order of zero crossings of resolved harmonics was set to 7. Distance thresholds for the acceptance or rejection of a segment were set to 5 samples (0.3125 ms) for resolved and 10 samples (0.625 ms) for unresolved harmonics. The *Signal to Noise Ratio (SNR)* values of the original mixture, the resulting signal after separation and those of the system by Hu and Wang [3] averaged over all speakers for a given intrusion are depicted in Tab. 1. 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 12%. In General the calculation of the zero crossing distances is less costly as the autocorrelation and zero crossings are physiologically more plausible. Furthermore the autocorrelation is physiologically rather implausible whereas a simple zero crossing detection and integration seems much more likely.

In summary, we showed that the zero crossing distances can be used efficiently as well to identify the pitch of multiple sound sources as to separate the sound sources.

References

- [1] B. C. J. Moore, *An introduction to the psychology of hearing*, Academic Press, London, 5th edition, 2003.
- [2] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filterbank," Tech. Rep., Apple Computer Co., 1993, Technical report #35.
- [3] 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.
- [4] A. de Cheveigne, "Pitch perception models," in *Pitch*, C. Plack and A. Oxenham, Eds. Springer Verlag, Cambridge, U.K., 2004.
- [5] Martin Cooke, *Modeling Auditory Processing and Organization*, Ph.D. thesis, Cambridge University, Cambridge, U.K., 1993.