Friedrich-Alexander-Universität
Erlangen-Nürnberg

Chair of Multimedia Communications and Signal Processing

Prof. Dr.-Ing. Walter Kellermann

Bachelor Thesis

Harmonic-Plus-Noise Model Pitch Estimation

Jutta Pirkl

Supervisor: Dipl.-Ing. Christian Hofmann

Erlangen, September 2013

Erklärung

Ich versichere, dass ich die vorliegende Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Erlangen, den 26. September 2013

Jutta Pirkl

Abstract

This thesis investigates a pitch detection algorithm named Least Squares Harmonic that is based on Harmonic-Plus-Noise modeling. The technique is claimed to be especially robust against noise, which is still a desired feature after more than 50 years of research.

A pitch frequency estimate for voiced speech periods is obtained as the fundamental frequency of the best-matching Harmonic-Plus-Noise model. For this purpose, the mean squared error between the modeled and the real speech signal for each possible frequency is minimized. The search interval of possible pitch frequencies can either be the whole human pitch range from 60 Hz to 400 Hz or a smaller interval around an initial estimate.

After preliminary experiments on the dependency of the algorithm on its input parameters, the algorithm's performance is evaluated by comparison with two classical pitch detection methods, namely the autocorrelation and cepstrum method, in terms of estimation accuracy, robustness against noise, and computation time. In the evaluation, two scenarios are examined. These are white Gaussian noise and wind noise environments.

By reference to the results of the experiments, the use of an initial estimate is recommended due to a higher risk for estimation errors, like multiple pitch errors, in the whole human pitch interval search. Furthermore, the algorithm is shown to be suitable to refine an initial estimate in white noise environments, especially at low signal-to-noise ratios. However, it performed inferior to the cepstrum method in wind noise and is very costly in terms of computation time owing to the extensive search over the set of possible pitch candidates.

Kurzfassung

Diese Arbeit untersucht den Least Squares Harmonic Algorithmus, einen Algorithmus zur Sprachgrundfrequenzschätzung, der auf einer Modellierung durch ein Harmonic-Plus-Noise Modell basiert. Die Methode gilt als besonders geeignet für verrauschte Signale, was trotz über 50 Jahren Forschung immer noch eine erwünschte Eigenschaft darstellt.

Den Schätzwert für die Sprachgrundfrequenz (auch Pitchfrequenz genannt) erhält man als Grundfrequenz des Harmonic-Plus-Noise Modells, das am besten an das originale Signal angepasst werden kann. Hierfür wird für jede mögliche Frequenz der mittlere quadratische Fehler zwischen dem modellierten und dem originalen Sprachsignal minimiert. Als Suchintervall für potientielle Pitchfrequenzen kann entweder der komplette Bereich menschlicher Pitchfrequenzen bei Gesprächen (60 Hz bis 400 Hz) oder ein kleineres Intervall um das Ergebnis eines Vorschätzers herum herangezogen werden.

Im Anschluss an vorläufige Experimente über die Abhängigkeit des Algorithmus von seinen Eingangsparametern wird die Methode durch einen Vergleich mit zwei klassischen Verfahren, der Autokorrelations- und der Cepstrummethode, in Bezug auf Schätzgenauigkeit, Robustheit gegen Rauschen und Rechenzeit evaluiert. Dabei werden die beiden Szenarien weißes gauß'sches Rauschen und Windrauschverhältnisse untersucht.

Auf Basis der Ergebnisse wird die Variante des Verfahrens mit Vorschätzer empfohlen, da die Suche über das komplette menschliche Pitchintervall ein größeres Risiko für Schätzfehler, wie beispielsweise Pitchvervielfachungen, birgt. Des Weiteren wird ge-

zeigt, dass die Methode dafür geeignet ist einen Vorschätzer in weißen Rauschverhältnissen zu verfeinern, was besonders bei sehr niedrigem Signal-zu-Rausch-Verhältnis deutlich wird. Allerdings ist sie bei Windrauschen der Cepstrummethode unterlegen und außerdem durch die umfangreiche Suche über alle möglichen Sprachgrundfrequenzen sehr rechenintensiv.

Contents

Chapter 1

Introduction

Many areas of speech signal processing can benefit from information about the fundamental frequency (pitch). Since it is an important parameter for voiced periods of speech, Pitch Detection Algorithms (PDAs) are integrated in a variety of applications, such as speaker recognition, speech synthesis, speech coding, speech enhancement, and many more.

Pitch estimation aims at providing a measurement of the oscillation frequency of the vocals folds during voiced speech periods that causes a quasi-periodic pattern in the time-domain signal. Actually, the term pitch denotes the subjectively perceived fundamental frequency of a speech sound, but it is mostly used with the same meaning as the physical fundamental frequency observable in the speech waveform [8].

Humans have the natural ability to distinguish between female and male speakers and even recognize known persons according to their voices. In this process, the pitch frequency plays an important part and can hence be used in automatic speaker recognition [21]. Another application area is speech synthesis, where the speech models must

be trained beforehand, based on pitch analyses, to learn the intonation of a particular human speaker. A prominent example related to speech synthesis is the physician Stephen Hawking, who is communicating via a computer with a voice synthesizer after having lost the ability to speak during a severe illness [7]. The possibility to synthesize speech with a small set of parameters, like a voiced/unvoiced discrimination and the fundamental frequency, is also utilized in speech coding to transmit a speech signal only requiring a low data rate [14]. In speech enhancement, undesired noise components in voiced speech can be eliminated with the help of the pitch frequency. For this purpose, the noisy signal is filtered with a comb filter, for example, that suppresses all frequency parts except the fundamental frequency and its harmonics. In this context, it is especially important to obtain very accurate pitch estimates, because deviations of the estimated from the true fundamental frequency multiply in the harmonics and thereby immediately lead to a significant suppression of desired speech components, whereas the noise residual increases.

As a result of these various applications, a large number of techniques for pitch detection has been developed over the last 50 years. However, their performance is varying dependent on the speaker's gender, the application and environmental condition. Whereas they mostly work very well in quiet environments by now, the performance rapidly drops at low Signal-to-Noise Ratios (SNRs) [17]. In practice, pitch detectors usually face noisy environments, though.

A method named Least Squares Harmonic (LSH), that is based on Harmonic-Plus-Noise (H+N) modeling and said to be especially robust against noise, is proposed by Abu-Shikhah et al. [1]. The purpose of this project is to implement and evaluate this pitch estimation technique. To work out the strengths and weaknesses, the LSH algorithm is compared with two conventional pitch detectors, the Autocorrelation Method (AUTOCOR) and Cepstrum Method (CEP), in different scenarios.

The thesis is organized in the following way: Chapter 2 begins by providing background knowledge about speech production and the corresponding source-filter model used in digital audio signal processing and some information about the pitch frequency. In Chapter 3, the principles and implementation of the LSH algorithm and the two classical PDAs are presented. Finally, in Chapter 4 the LSH method is evaluated in different scenarios and the experimental results are compared to the AUTOCOR and CEP methods.

Chapter 2

Fundamentals of Speech Production

Before presenting the theory and implementation of the PDAs, it is necessary to address the fundamentals of the speech production model that will serve as the basis for the pitch tracking techniques.

This so-called source-filter model [22, 19] assumes a speech signal to be the output of a time-varying linear system representing the effects of the vocal tract, as depicted in Figure 2.1. The energy for speech production is provided by the sub-glottal system comprised of the lungs, bronchi and trachea. Vocal tract denotes the area from the opening between the vocal folds (glottis) to the mouth (oral cavity) or nose, respectively. It can be thought of as a concatenation of acoustic tubes. When the lungs contract, air flows through the vocal tract where it is shaped due to the frequency selectivity of these tubes and then radiated at the lips.

Speech sounds may be divided into three main classes according to the excitation mode and the state of the vocal cords respectively, which are: voiced, unvoiced and plosive sounds [22]. *Voiced* sounds are generated by oscillating vocal folds modulating

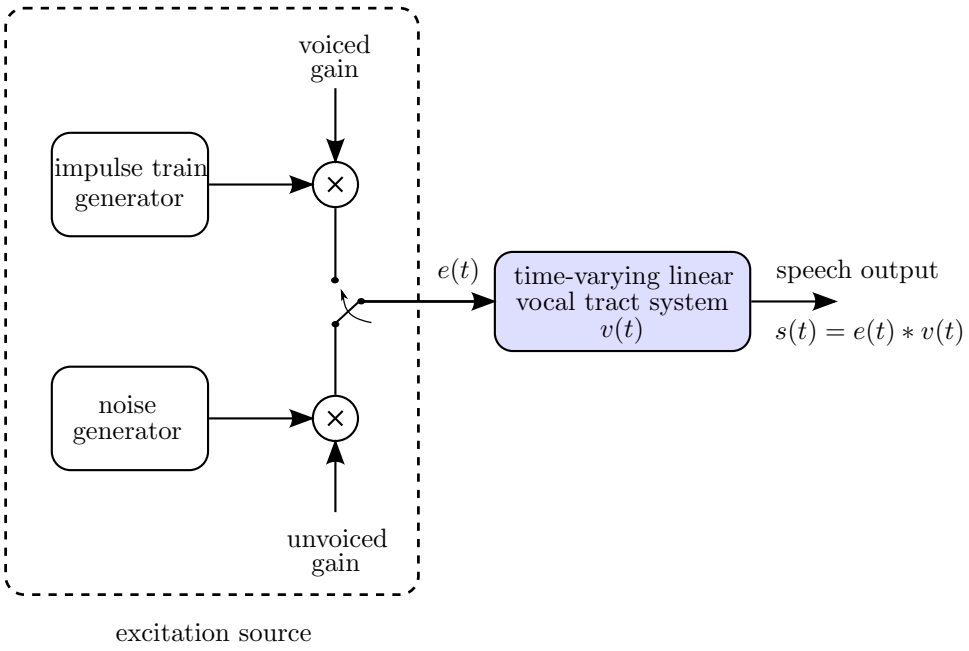


Figure 2.1: Source-filter model of speech production [26].

the air flow from the lungs. In this way, quasi-periodic pulses of air are released to excite the vocal tract. To take an example, all vowels or the consonant /m/ are of this type. The speaker can feel the oscillation of the vocal cords when touching the larynx. On the other hand, during *unvoiced* sounds, such as /f/ and /s/, the vocal folds are open. They are produced by narrowings in the vocal tract (mostly in the oral cavity) causing the air flow to become turbulent and thereby resulting in a noise-like excitation [8]. For the production of *plosive* sounds, like /t/, the air flow from the lungs builds up pressure behind a closing above the glottis and is then abruptly released.

To sum up, the excitation source $e(t)$ in the speech production model is either a glottal impulse train with the period according to the oscillation of the glottis or a randomly distributed noise generator, as represented by the switch in Figure 2.1. Thus, the speech output can be computed as the convolution of $e(t)$ with the vocal tract impulse response $v(t)$. Note that the excitation mode and the parameters of the vocal tract change over time. As they change relatively slowly, the process can be

called quasi-stationary and the properties are assumed to be constant for periods of about 10 ms. For this reason, the speech signal must be examined in a short-term analysis, where the signal is divided into short, usually overlapping segments that are analyzed separately [22].

The time-varying spectral characteristics can be observed in a spectrogram that displays the frequency content in a time-frequency plane [9]. One example is given in Figure 2.2. The colors in this spectrogram illustrate the logarithmic energy at a particular time segment and frequency, where red means high energy. Unvoiced parts show a broad noise-like distribution of the frequency content, whereas voiced regions are characterized by the harmonic structure that is caused by the periodicity of the time-domain signal and can be seen by the dominant, parallel horizontal lines in the spectrogram. The fundamental frequency is referred to as the pitch frequency or simply pitch. It is the reciprocal of the pitch period, illustrated in Figure 2.3.

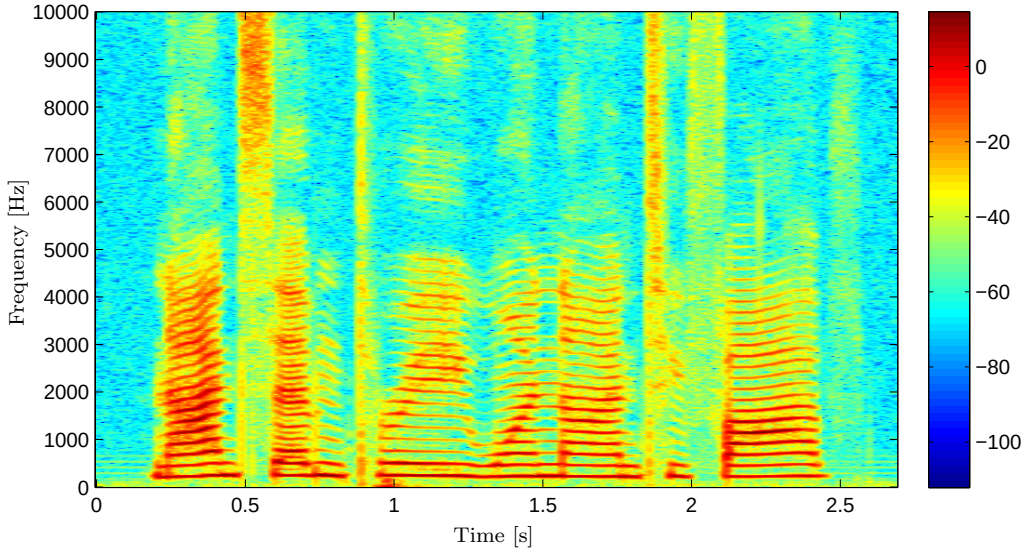


Figure 2.2: Spectrogram of the utterance ”9 7 3 1 9 2 5” spoken by a female and sampled at 20 kHz.

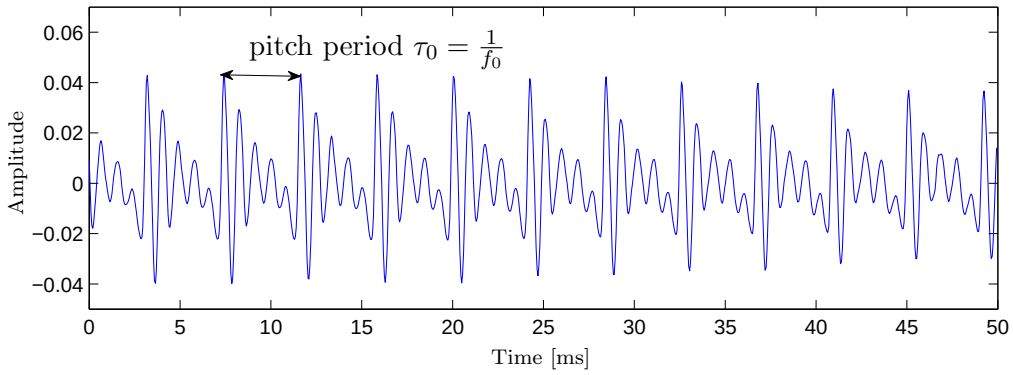


Figure 2.3: Voiced speech segment showing a periodic pattern according to the pitch period.

A human speaker varies within a certain range of pitch frequencies that is somewhere between 60 Hz and 400 Hz [19] and characteristic for this person. Males have a lower pitch (average: 120 Hz) than females (average: 240 Hz), which is the result from longer and more massive male vocal folds [8]. Variations of the pitch can be heard as the intonation. For instance, the pitch contour increases at the end of a question. Along with loudness and rhythm, it represents the so-called prosodic information, from which one can deduce different meanings of an expression or the emotion of a speaker.

Chapter 3

Pitch Detection Algorithms

As information about the pitch frequency is required in many areas of speech research, a variety of methods for pitch determination has been developed to date. They can generally be classified into three categories: time-domain, frequency-domain and hybrid (both time- and frequency-domain) pitch detectors. Whereas time-domain pitch detectors directly work with the speech waveform, the frequency-domain techniques utilize the harmonic characteristic in the speech spectra [21]. In order to compare the performance of the Least Squares Harmonic (LSH) algorithm, two of these pitch detection algorithms were additionally implemented, a frequency-domain PDA, the Cepstrum Method (CEP) and a time-domain PDA, the Autocorrelation Method (AUTOCOR). They were chosen because the Cepstrum Method is known as an especially accurate pitch extraction method [21] and the Autocorrelation Method due to its robustness against additive white Gaussian noise [24]. In this chapter, all three algorithms are presented.

3.1 Cepstrum Pitch Estimation

The Cepstrum method was already developed in the 1960s by A. M. Noll [15]. The basic principle of this method is to compute the cepstrum of the time frame being analyzed, because the cepstrum contains a strong peak at the location corresponding to the instantaneous pitch period if the frame is voiced. Hence, pitch determination reduces to peak detection in the cepstral domain.

3.1.1 Cepstrum

The cepstrum is defined as the square of the inverse Fourier transform of the logarithm power spectrum [15]. Thus, the cepstrum for a speech frame $s[k]$[1] can be calculated as

$$C[n] = IDFT\left\{ \log\left(\left|DFT\{s[k]\}\right|^2\right)\right\}^2, \tag{3.1}$$

where n is the so-called quefrency index. Due to the logarithm, the resulting domain is neither time-domain nor frequency-domain, but a new domain that is called the quefrency-domain (a "reversed" frequency) and has the unit seconds. In this thesis $C[n]$ is always a real-valued and even sequence since the magnitude-square spectrum of a real-valued time sequence is real and even [8].

The explanation for the possibility to use the cepstrum in pitch detection is the following: If a signal is quasi-periodic in time-domain, the frequency spectrum consists of equally spaced impulses located at the fundamental frequency and its harmonics and is hence itself periodic. The straightforward approach to determine this period is to take the Fourier transform of the power spectrum [15]. As the resulting peak is

[1]Throughout this thesis the square bracket indicates the discrete-time domain, k is the sample index.

broadened by the effects of the vocal tract filter (see Chapter 2), the Fourier transform is instead applied to the logarithm of the power spectrum. That way the effects of the excitation source and the vocal tract are separated [15], because

$$\log\left(|S(e^{j\Omega})|\right) = \log\left(|E(e^{j\Omega})| \cdot |V(e^{j\Omega})|\right) = \log\left(|E(e^{j\Omega})|\right) + \log\left(|V(e^{j\Omega})|\right) \tag{3.2}$$

leads to

$$C_s[n] = IDFT\left\{ \log\left(\left|S(e^{j\Omega})\right|^2\right)\right\}^2 = C_e[n] + C_v[n], \tag{3.3}$$

where $C_e[n]$ denotes the cepstrum of the periodic excitation source with peaks at high quefrencies, whereas the smooth envelope of the vocal tract filter contains mostly low-quefrency energy in $C_v[n]$.

This is illustrated in Figures 3.1 and 3.2. Figure 3.1 shows a short-time log power spectrum of a voiced speech frame with its typical harmonic structure that is recognizable by the periodic ripples. Information about the underlying fundamental frequency is captured by the location of the marked peak in the corresponding cepstrum, as depicted in Figure 3.2.

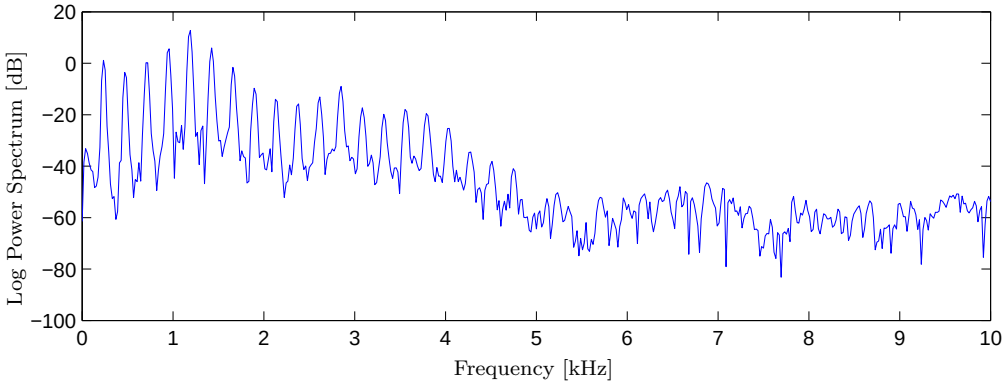


Figure 3.1: Log power spectrum of a voiced speech frame showing periodic ripples caused by the pitch frequency ($f_s = 20kHz$, $N_{FFT} = 1024$).

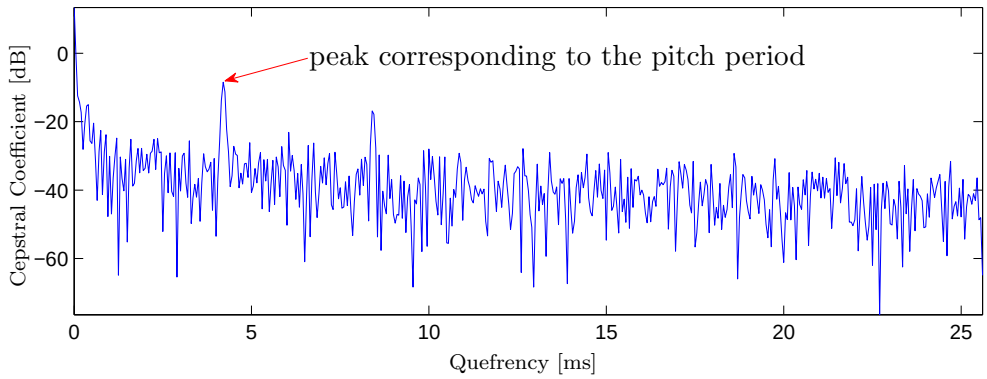


Figure 3.2: Cepstrum of a voiced speech frame ($f_s = 20kHz$, $N_{FFT} = 1024$).

3.1.2 Peak Detection

After the computation of the cepstrum, the correct peak has to be selected. For this purpose, the cepstrum is only examined in the interval [2.5 ms; 15 ms], since the pitch frequency is within the range of 60 Hz to 400 Hz. As the cepstral coefficients decrease with increasing quefrency, the interval is multiplied with a linearly increasing weight from 1 to 3 [15, 2]. Additionally, the sensitivity in the vicinity of the pitch period of the previous frame is increased by weighting this area with a von Hann window [9] with offset one. Then the maximum peak is determined. Finally, the pitch period of the frame is obtained as the center of gravity of the maximum peak and its previous and following quefrency index [6].

3.2 Autocorrelation Pitch Estimation

The second PDA being implemented is the Autocorrelation method [20, 24, 11].

From an Autocorrelation Function (ACF), one can deduce the similarity of a signal and its delayed versions, i.e. it shows a peak whenever the signal is similar to its delayed

version. The ACF of a real-valued speech section $s[k]$ is estimated by

$$\varphi_{ss}[\kappa] = \frac{1}{N} \sum_{k=0}^{N-\kappa-1} s[k+\kappa]s[k] \qquad for \;\; \kappa \geq 0, \tag{3.4}$$

where κ is the lag number and N is the length of $s[k]$.

As a voiced speech frame is quasi-periodic in time-domain with period τ_0, its ACF consequently has peaks at lags equal to integer multiples of τ_0. The largest peak can obviously be found at lag $\kappa = 0$. When using the biased ACF estimator of Equation (3.4), the other peaks decrease in amplitude with increasing lags due to the frames' limited duration. Therefore, the estimate for the pitch period is obtained by determining the location of the second peak [24] (see example in Figure 3.3). In the implementation the interval [2.5 ms; 15 ms] is searched for the maximum peak, as in the CEP method.

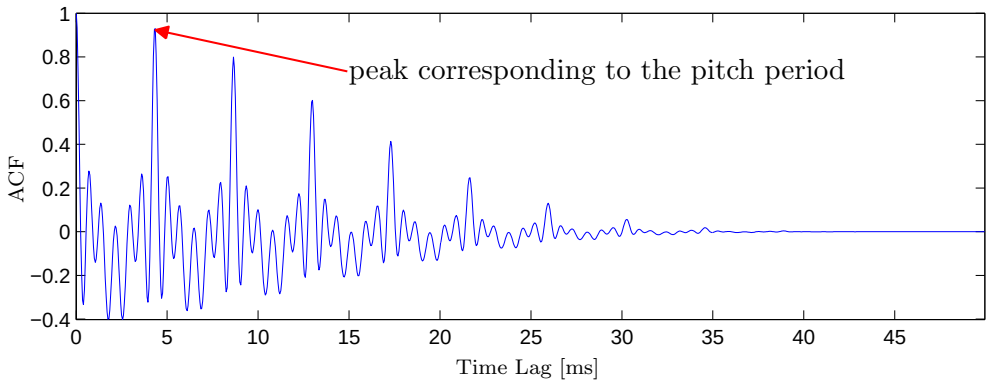


Figure 3.3: Autocorrelation function of a voiced speech frame normalized with the signal energy ($f_s = 20kHz$).

3.3 Harmonic-Plus-Noise Model Pitch Estimation

Turning now to the technique for pitch estimation being in the focus of this thesis, the H+N model pitch estimation, this section first describes the H+N model and based

on that the derivation of the LSH algorithm.

3.3.1 Harmonic-Plus-Noise Model

The H+N model is a special case of the sinusoidal representation for speech signals developed by McAulay et al. [13]. This representation is obtained by the help of the speech production model. As previously described in Chapter 2, the speech production model assumes speech to be the output of a linear time-varying vocal tract filter that is excited by a glottal signal generator. In the sinusoidal model a sum of sine waves with arbitrary amplitudes, frequencies and phases is used for the excitation signal.

After passing it through the vocal tract filter the output, i.e. the discrete-time representation of a short speech segment $s[k]$, can be written as

$$s[k] = \sum_{i=1}^{M} C_i \cos\left[2\pi \frac{f_i}{f_s} k + \phi_i\right] = \sum_{i=1}^{M} C_i \cos\left[\omega_i k + \phi_i\right], \tag{3.5}$$

where M is the number of sinusoids and C_i, f_i and ϕ_i their corresponding amplitudes, frequencies and phases, which can be assumed to be constant over the duration of the time frame due to the quasi-stationarity of speech [1]. The normalized frequency is denoted by $\omega = 2\pi \frac{f}{f_s}$.

The H+N model utilizes the assumption that voiced speech is harmonic and thereby decomposes $s[k]$ into two parts: the harmonic component $h[k]$ (quasi-periodic), representing the voiced speech part and the noise component $n[k]$ (non-periodic), representing the unvoiced speech part as well as additional other noise. In the harmonic component the first P sine waves in Equation (3.5) with frequencies being multiples of the fundamental frequency f_0 are summarized. The remaining sinusoids are summed

up to the noise component. Consequently, the H+N model is given by [1]:

$$s[k] = h[k] + n[k] \tag{3.6a}$$

$$= \sum_{i=1}^{P} C_i \cos\left[2\pi i \frac{f_0}{f_s} k + \phi_i\right] + n[k] \tag{3.6b}$$

$$= \sum_{i=1}^{P} C_i \cos\left[i\omega_0 k + \phi_i\right] + n[k]. \tag{3.6c}$$

Please note that all considerations in this chapter are for a short time frame and therefore the fundamental frequency can also be assumed time-invariant over the duration of this frame.

3.3.2 Least Squares Harmonic Algorithm

The pitch detection algorithm named LSH [1] can now be derived on the basis of the presented H+N model.

First of all, $h[k]$ in Equation (3.6c) is rearranged by applying one of the addition theorems [4, Eq. 2.91] to split up the cosine term:

$$h[k] = \sum_{i=1}^{P} C_i(\cos\left[i\omega_0 k\right]\cos\left[\phi_i\right] - \sin\left[i\omega_0 k\right]\sin\left[\phi_i\right]) \tag{3.7a}$$

$$= \sum_{i=1}^{P} A_i \cos\left[i\omega_0 k\right] - B_i \sin\left[i\omega_0 k\right], \tag{3.7b}$$

where $A_i = C_i \cos\left[\phi_i\right]$ and $B_i = C_i \sin\left[\phi_i\right]$.

In order to estimate the fundamental frequency and the amplitudes and phases of the model such that they approximate the real speech signal as good as possible, they are adapted to the audio recordings according to the Minimum Mean Squared Error (MMSE) criterion, as this assures robustness against additive white Gaussian noise [14].

The Mean Squared Error (MSE) between the speech frame $s[k]$ of length N and the harmonic part $h[k]$

$$MSE = \frac{1}{N}\sum_{k=0}^{N-1}(s[k] - h[k])^2 \tag{3.8a}$$

$$= \frac{1}{N}\sum_{k=0}^{N-1}(s[k] - \sum_{i=1}^{P} A_i \cos[i\omega_0 k] - B_i \sin[i\omega_0 k])^2 \tag{3.8b}$$

is minimized for a given fundamental frequency f_0 and ω_0 respectively, by setting the partial derivatives $\frac{\partial MSE}{\partial A_j}$ and $\frac{\partial MSE}{\partial B_j}$ to zero:

$$\frac{\partial MSE}{\partial A_j} = 0 \quad and \quad \frac{\partial MSE}{\partial B_j} = 0, \quad for \;\; j = 1, 2, ..., P. \tag{3.9}$$

Thus resulting in $2P$ linear equations. Therefore, it is more convenient to switch to a matrix notation at this point and summarize P equations each:

$$\mathbf{Y_1} = \mathbf{QA} + \mathbf{RB} \tag{3.10}$$

$$\mathbf{Y_2} = \mathbf{SA} + \mathbf{TB}, \tag{3.11}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the vectors of the unknowns A_j and B_j with size $P \times 1$ and the remaining matrices are defined as:

$$\mathbf{Q}[i,j] = \sum_k \cos[i\omega_0 k]\cos[j\omega_0 k] \tag{3.12a}$$

$$\mathbf{R}[i,j] = -\sum_k \sin[i\omega_0 k]\cos[j\omega_0 k] \tag{3.12b}$$

$$\mathbf{S}[i,j] = \sum_k \cos[i\omega_0 k]\sin[j\omega_0 k] \tag{3.12c}$$

$$\mathbf{T}[i,j] = -\sum_k \sin[i\omega_0 k]\sin[j\omega_0 k] \tag{3.12d}$$

$$\mathbf{Y_1}[i,j] = \sum_k s[k]\cos[j\omega_0 k] \tag{3.12e}$$

$$\mathbf{Y_2}[i,j] = \sum_k s[k]\sin[j\omega_0 k], \tag{3.12f}$$

where $i, j = 1, 2, ..., P$ and $k = 0, 1, ..., N-1$.

Solving Equations (3.10) and (3.11) for $\mathbf{A}$ and $\mathbf{B}$ leads to the equations

$$\mathbf{A} = (\mathbf{S} - \mathbf{T}\mathbf{R}^{-1}\mathbf{Q})^{-1}(\mathbf{Y_2} - \mathbf{T}\mathbf{R}^{-1}\mathbf{Y_1}) \tag{3.13}$$

$$\mathbf{B} = \mathbf{R}^{-1}(\mathbf{Y_1} - \mathbf{Q}\mathbf{A}) \tag{3.14}$$

and with that the amplitudes C_i and phases ϕ_i of the sine waves in the harmonic component for the predefined fundamental frequency can be determined [1].

Unfortunately, there is no closed-form analytical solution for the fundamental frequency, therefore the matrices defined in Equation (3.12) and Equations (3.13) and (3.14) have to be calculated repeatedly for a range of possible fundamental frequencies. The final solution for f_0 is the one resulting in the smallest MSE. Hence, the LSH is very costly in terms of computation time.

To sum up, the LSH algorithm can be divided into the following steps [1, 18]:

1. Define a search interval vector of possible pitch frequencies $f = [f_0, f_0+\delta, ..., f_{0,max}]$, where δ is the step size with which the search interval is sampled.

2. Repeat for each element in the search interval vector starting with the lower interval boundary until all elements are processed:

   2.1. Compute the matrices $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{S}$, $\mathbf{T}$, $\mathbf{Y_1}$ and $\mathbf{Y_2}$ according to Equation (3.12).

   2.2. Solve for the unknowns $\mathbf{A}$ and $\mathbf{B}$.

   2.3. Calculate the estimated harmonic component using Equation (3.7b).

   2.4. Compute and store the MSE between the speech frame and the estimated harmonic component.

3. Designate the frequency resulting in the smallest MSE as the pitch frequency for the frame.

This procedure is also illustrated in Figure 3.4. To obtain a pitch contour for a whole speech signal the process has to be performed iteratively over all frames which the signal is divided into.

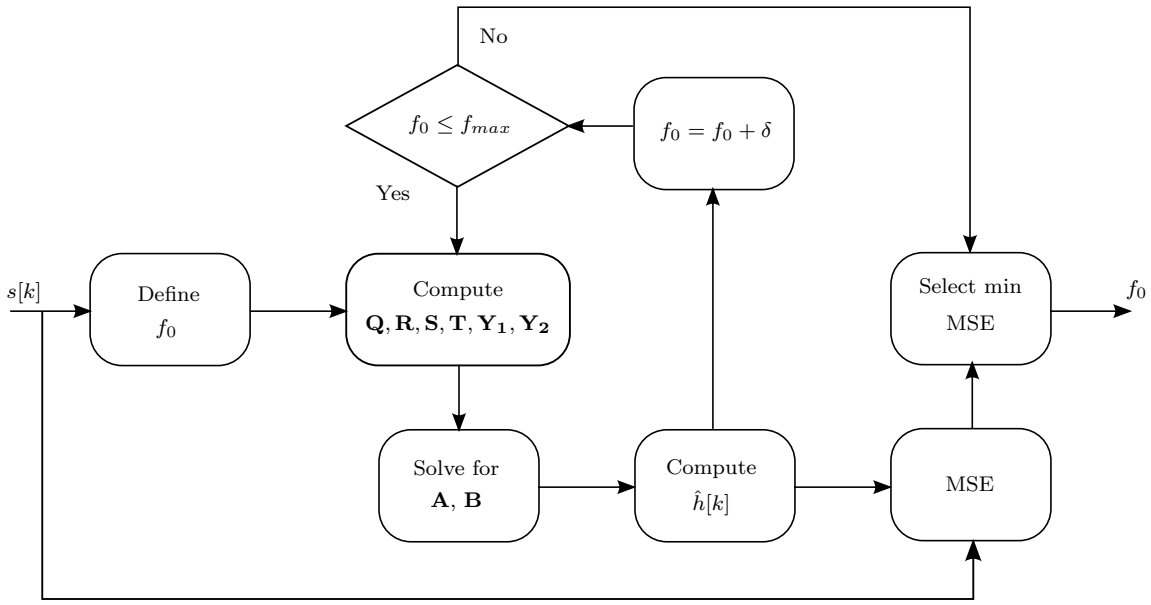


Figure 3.4: Flowchart LSH Algorithm [1].

In this thesis, three ways for the definition of the search interval are examined:

a) *LSH Version Complete Interval Fine (LSH CIF)*: The search interval vector is the whole range of potential human pitch frequencies from 60 Hz to 400 Hz with a small step size δ (e.g. $\delta_{fine} = 0.1$ Hz).

b) *LSH Version First Rough Then Fine (LSH FRTF)*: Firstly, the whole pitch range is searched with a bigger step size (e.g. $\delta_{rough} = 5$ Hz) and secondly, the rough estimate is refined with a smaller search interval and finer step size around it.

c) *LSH Version Initial Estimate (LSH IE)*: The search interval is obtained by an initial estimate from another PDA as a small, but accurate search interval around this estimate.

Furthermore, in the implementation in step 2.2 the following equations are used to solve for the unknowns $\mathbf{A}$ and $\mathbf{B}$

$$\mathbf{A} = (\mathbf{Q} - \mathbf{R}\mathbf{T}^{-\mathbf{1}}\mathbf{S})^{-\mathbf{1}}(\mathbf{Y_1} - \mathbf{R}\mathbf{T}^{-\mathbf{1}}\mathbf{Y_2}) \tag{3.15}$$

$$\mathbf{B} = \mathbf{T}^{-\mathbf{1}}(\mathbf{Y_2} - \mathbf{S}\mathbf{A}), \tag{3.16}$$

which are the result of solving Equation (3.11) for $\mathbf{B}$ and subsequently inserting $\mathbf{B}$ in Equation (3.10) instead of the other way round, as it is done in [1]. The reason for this is the ill-conditioning of matrix $\mathbf{R}$, which caused serious problems regarding the performance of the LSH. If a matrix is ill-conditioned, it is close to singular and the computation of its inverse or the solution to a set of linear equations respectively is subject to numerical errors [12]. In contrast, matrix $\mathbf{T}$ has shown to be considerably better in terms of well-conditioning.

The algorithms are implemented in MATLAB and integrated in subclasses which inherit from a general super class for pitch estimation to benefit from enclosure and exploit similarities. The corresponding Unified Modeling Language (UML) diagram can be found in Appendix B.

Chapter 4

Objective Evaluation

In this chapter, the H+N model pitch detector is evaluated using objective performance measures that are, along with other information about the experimental setup, described first. Then preliminary tests to investigate the performance dependency on the input parameters of the LSH algorithm are presented and finally, its performance is compared with the two classical PDAs, the AUTOCOR and CEP methods, in different scenarios. The results of the experiments are discussed with various charts, where the underlying exact numbers can be found in Appendix A in case of interest.

4.1 Experimental Setup

4.1.1 Database

The performance of the three PDAs is evaluated with the database from Keele University developed by Plante et al. [17]. It contains ten speech signals with durations of 27 to 40 seconds from five mature male and five mature female speakers (labeled M1 - M5 and F1 - F5) reading a phonetically balanced text. The data was recorded in a soundproof room and sampled at 20 kHz with a 16 bit resolution.

In order to examine the accuracy of the pitch estimates, a reference pitch contour for each speech file is required. In the experiments, a corrected version of the reference files for the Keele Database created by Frederico Flego [5] was used, because it had shown to be more appropriate in a visual inspection in the spectrograms than the original version. The reference pitch contours also include a voiced/unvoiced discrimination that was adopted directly, since the task of making a voiced/unvoiced decision is not a concern in this work. Pitch frequency references in case of voiced excitation or a tag for unvoiced excitation are provided every millisecond. To be able to compare with the reference data and compute an objective deviation measure, which will be described in the next section, the pitch frequency estimate from the PDAs for each analysis frame is assigned to its center.

4.1.2 Performance Measures

To quantitatively measure the accuracy of the estimation, an objective error measure, the Weighted Gross Pitch Error (GPE), is calculated for every estimation. It is defined as [2]

$$GPE = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{E_n}{E_{max}} \right)^{0.5} \left| \frac{\hat{f}_n - f_n}{f_n} \right|, \tag{4.1}$$

where N is the number of voiced frames, E_n the short-time energy of the n-th voiced frame, and E_{max} the maximum short-time frame energy of the whole signal. Thus, the GPE expresses the average weighted relative deviation of the estimated $\hat{f}_n$ from the reference pitch frequencies f_n. The energy weighting is applied because the perceived distortion due to pitch errors is becoming less significant with decreasing frame energy [27].

Another error parameter is the Multiple Pitch Error Rate (MPER). It considers the common problem in pitch estimation that multiples, like twice the true fundamental frequency, are mistakenly chosen as pitch frequency. In this work, the MPER indicates

the percentage of voiced frames with estimates being twice, three times, a half, or a third of the true pitch.

4.1.3 Experimental Details

All experiments are conducted in the same general way, only with varying input signals. After the speech samples are stored in a buffer, the pitch estimation algorithm is performed block by block. Each frame is weighted with a von Hann window to profit from its good sidelobe structure [9]. Once the pitch contour is complete, a median filter of order five is applied to remove single or two successive isolated errors [2]. Finally, the frames are classified voiced or unvoiced according to the reference file and the GPE of the voiced estimates is computed. The specific settings are provided in Table 4.1.

The window size is an important factor in pitch detection that is subject to the trade-off between detecting rapid changes, where short frames are required, and the ability to accurately determine the periodicity, where longer frames are of advantage. In this project, it was chosen 50 ms, to ensure a length of at least three times the pitch period for all pitch frequencies greater than 60 Hz [20]. The window length was confirmed to be necessary in a test that showed a 5% decrease in the average GPE for all three PDAs in comparison to a 20 ms window. This improvement mainly stems from male speakers, usually having low pitch frequencies and longer pitch periods, respectively. Hence, 33672 frames in total were analyzed in each experiment.

In the evaluations until Section 4.3.2, the initial estimates for the LSH IE version are obtained from the AUTOCOR method, because of its low complexity and robustness against white Gaussian noise, as will be seen later. The search interval width was determined according to the average 95%-quantile of the absolute pitch deviation of the AUTOCOR method for clean speech signals (approx. 10 Hz) and an additional 10 Hz buffer taking into account worse performance in noisy environments. In other words, the search interval allows an absolute deviation of the initial estimate from the

true pitch frequency of 20 Hz to still contain the true pitch.

Table 4.1: Experimental settings.

General	
window size:	50 ms
frame shift:	10 ms
AUTOCOR	
see general	
CEP	
FFT-length:	1024
LSH	
number of harmonics P:	15
LSH IE	
search interval width:	40 Hz (20 Hz positive and negative deviation from the initial estimate)
step size δ:	0.1 Hz
LSH FRTF	
search interval:	[60 Hz; 400 Hz]
step size δ_{fine}:	0.1 Hz
step size δ_{rough}:	5 Hz

4.2 Preliminary Investigations on the LSH Algorithm

Before comparing the LSH algorithm to the CEP and AUTOCOR methods, this section deals with the best choice for the input parameters of the LSH pitch tracker. These are the number of harmonics P and the definition of the search interval.

4.2.1 Dependency on the Number of Harmonics

To examine the influence of the number of harmonics P on the accuracy of the pitch extraction, experiments (procedure as described in Section 4.1.3) were conducted on the clean speech signals for $P = \{5, 10, 15, 20\}$. One might expect the estimation to be most accurate when P resembles the true number of harmonics in the signal. Since all signals are sampled at 20 kHz, the maximum frequency would be 10 kHz, but by inspection of several spectrograms, like the one in Chapter 2, the clearly visible harmonics only reach up to about 4-5 kHz. For an average female speaker this results in about 20 harmonics, for male speakers the number of harmonics is greater because of their lower pitch. Starting from $P = 25$, the matrix $\mathbf{T}$ also shows ill-conditioning[1]. Therefore, greater values for P were not examined.

The simulation was run for the LSH versions LSH FRTF and LSH IE using the AUTOCOR method as initial estimate. Version LSH CIF cost too much computation time, but will be mentioned in the following subsection about the definition of the search interval.

The results are presented graphically in Figures 4.1 and 4.2 and the exact data for the graphs can be found in the appendix (Table A.1). Figure 4.1 shows the results obtained with the LSH version *LSH IE AUTOCOR*. From this figure, it can be

[1]Note that matrix $\mathbf{T}$ depends on f_s, for other sampling rates the ill-conditioning behaves differently. For example, at $f_s = 8$ kHz the ill-conditioning of $\mathbf{T}$ is worse, i.e. more matrices are affected

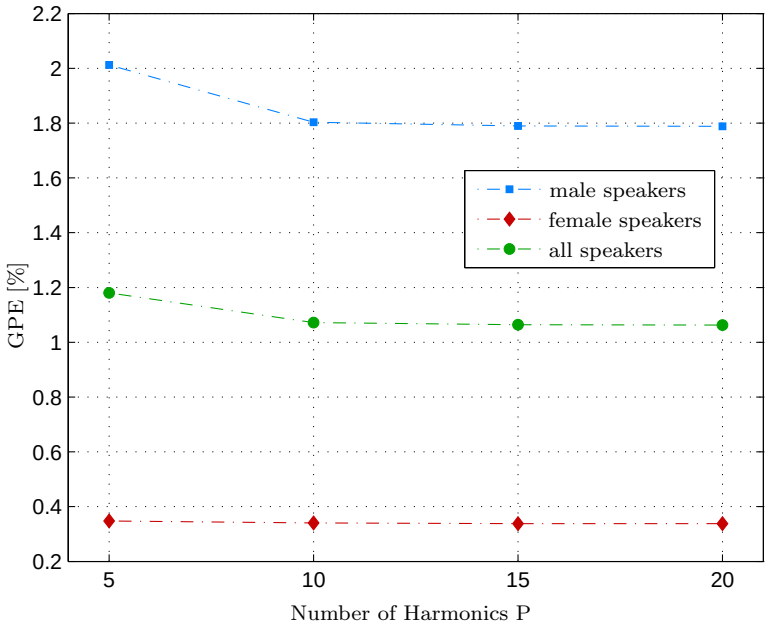


Figure 4.1: Dependency of the LSH IE AUTOCOR algorithm on the number of harmonics.

observed that the average GPE for all speakers shows a tendency to fall with increasing P. However, the decrease stops at $P = 15$ and nearly no improvements can be achieved anymore despite additional computational cost. The GPE for female speakers is saturated earlier, being almost constant over P in contrast to the GPE for male speakers, which is probably caused by the smaller amount of harmonics in the real signal. Furthermore, the estimation is significantly more accurate for female speech. This behavior is inherited from the AUTOCOR method (see Section 4.3). Based on these results, the following experiments will use $P = 15$ to achieve maximum accuracy.

Contrary to expectations, the total GPE of the *LSH FRTF* (depicted in Figure 4.2) increases after a primary drop from $P = 5$ to $P = 10$. A significant contribution to this is the increasing occurrence of multiple pitch errors, as shown in Table 4.2. The LSH IE version cannot have this problem, because the small interval around the initial estimate does not include any pitch multiples. Hence, the MPER is constant over P, originating from the AUTOCOR method itself. In contrast to the LSH IE AUTOCOR, the LSH FRTF version generally seems to be better for male speakers.

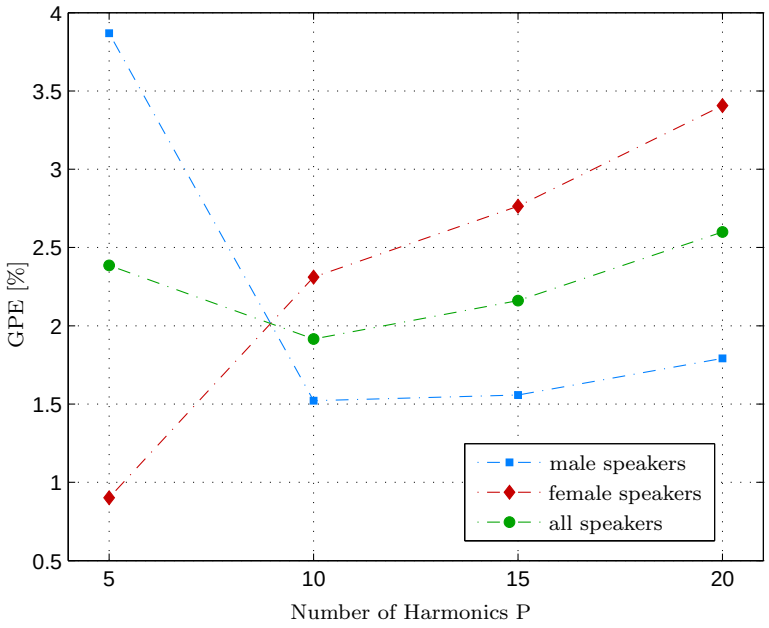
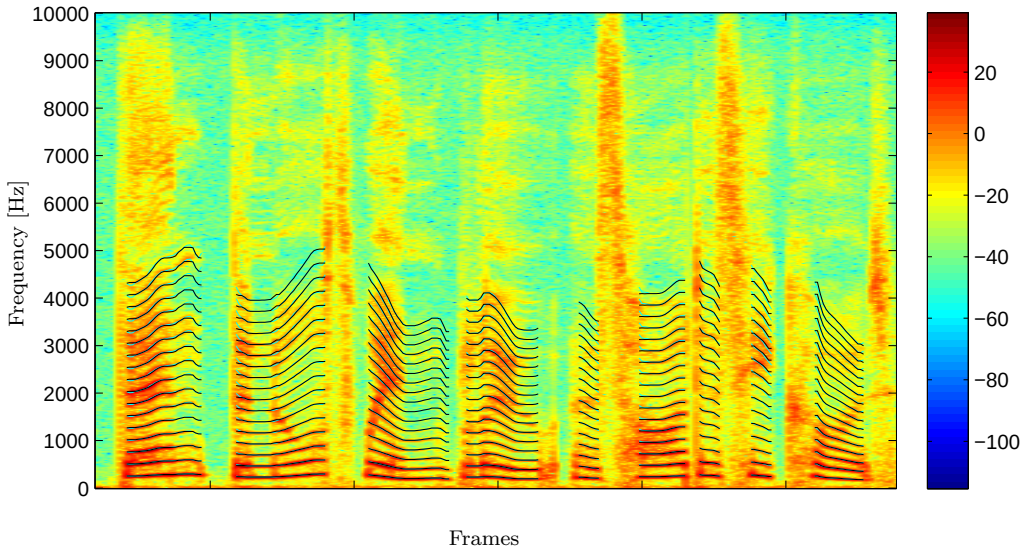


Figure 4.2: Dependency of the LSH FRTF algorithm on the number of harmonics.

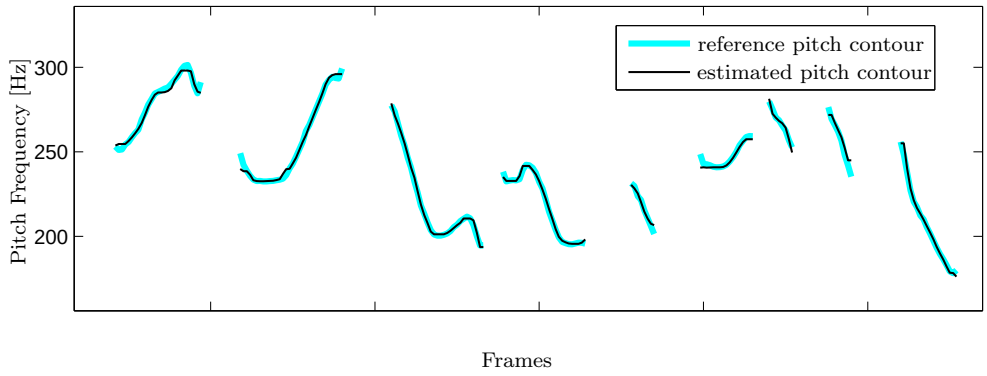
Table 4.2: Multiple pitch error rate for different numbers of harmonics.

MPER in %	P = 5	P = 10	P = 15	P = 20
LSH IE AUTOCOR	2.2	2.2	2.2	2.2
LSH FRTF	7.5	10.4	11.2	13.0

All in all, the GPE for both versions varies on a low level suggesting the LSH to be competitive in accuracy with the other two conventional PDAs. This is also clearly supported in plots of the spectrograms with the estimated pitch contours. An example from the experiments is provided in Figure 4.3.



(a) Section of the spectrogram with the estimated pitch contour and its harmonics.



(b) Comparison of the estimated and reference pitch contour.

Figure 4.3: Representative example of an estimation with the LSH IE AUTOCOR version (speaker F5).

4.2.2 Definition of the Search Interval

Based on the results of the previous subsection, the LSH IE version seems to be superior to the LSH FRTF, because the LSH FRTF struggles a lot with multiple pitch errors.

To examine this problem, another test was carried out, comparing the LSH FRTF pitch contour of a signal with extremely high MPER (20,3%) to one obtained from searching the whole human pitch range immediately with the fine step size (LSH CIF version) instead of a preceding rough search. However, no improvement was accomplished at a high cost of computation time. The MPER even rose to 36,1%. Most estimates in the contour were identical, differences were interestingly nearly always multiples of each other. Therefore, the LSH CIF version was not examined further.

4.3 Performance Comparison

Having conducted preliminary experiments on the LSH algorithm, the final section of this chapter compares the proposed method with the two conventional PDAs, the AUTOCOR and CEP pitch detectors, with respect to the following performance criteria: estimation accuracy, robustness against noise, and computation time. For this purpose, two different scenarios are examined. Firstly, white Gaussian noise is added to the speech signals and secondly, wind noise conditions are simulated.

As described in the previous section, the LSH FRTF struggles with multiple pitch errors. This is a common problem in pitch detection and there are possibilities to eliminate them. For this reason, an additional separate consideration, where all frames containing such errors were not included in the GPE calculation, is made.

4.3.1 Robustness against White Gaussian Noise

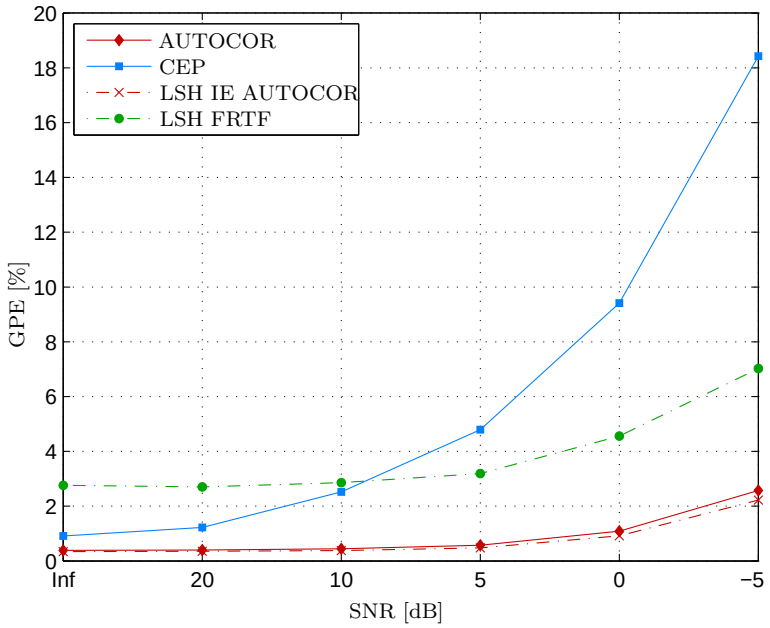
In the first scenario, white Gaussian noise at different SNR levels is added to the Keele speech files. These are $SNR = \{Inf,\ 20,\ 10,\ 5,\ 0,\ -5\}$ dB, where Inf dB represents the original speech signals without additional noise.

The results obtained from the simulations are shown in Figures 4.4 to 4.6. Exact numbers for the charts are provided in Table A.3. It is apparent from all three figures that the GPE of each pitch tracker is, as expected, increasing with decreasing SNR. Around 5 dB, the rise becomes steeper. Furthermore, the methods perform differently for male and female speech.

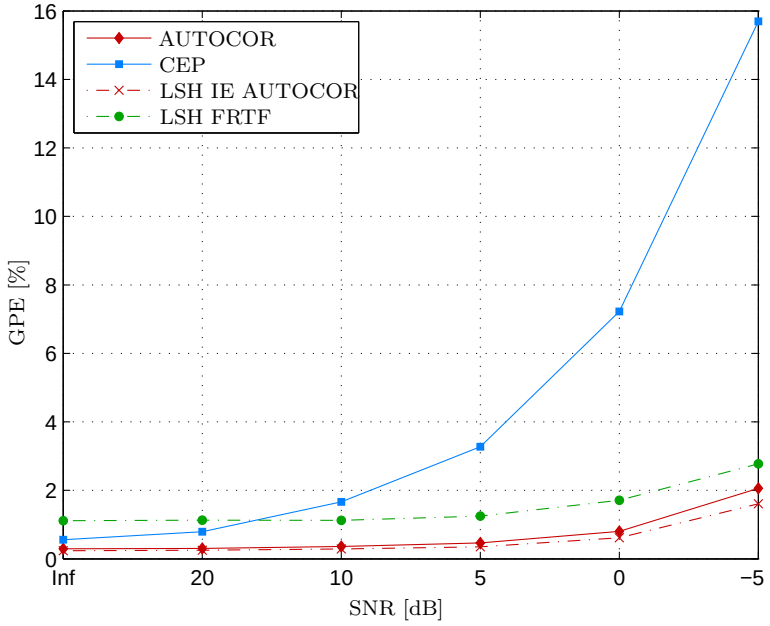
For female speakers, as depicted in Figure 4.4a, the LSH IE using the AUTOCOR method to obtain the search interval performs best for all SNRs, exploiting the accurate initial estimate of the AUTOCOR method. Thus, the preceding detection can always be refined. The independent version LSH FRTF is inferior to the other PDAs for clean and mildly noisy signals, but outperforms the CEP method at low SNR levels starting from about 10 dB. While the GPE of the CEP is close to the ones of the AUTOCOR and LSH IE AUTOCOR for clean speech signals, it rises more steeply with increasing noise level, seeming not to be robust against white Gaussian noise.

Figure 4.4b shows the results of the GPE calculation without frames containing multiple pitch errors. From this figure, one can see that the curve for the LSH FRTF has significantly shifted downward, already outperforming the CEP method at a higher SNR level and being closer to the AUTOCOR and LSH IE AUTOCOR methods than before. However, it stays inferior to them.

For male speakers (Figure 4.5a), the CEP and LSH FRTF are competing for the best performance. They seem to be more accurate for male than female speech signals. Whereas the CEP method has the lowest GPE for clean signals, it is outperformed by the LSH FRTF between 10 dB and 0 dB because of the CEP's steeper rise of the GPE.



(a) Dependency of the average GPE on different SNR levels.



(b) Dependency of the average GPE on different SNR levels without frames containing multiple pitch errors.

Figure 4.4: Performance of the PDAs at different SNR levels for *female speakers*.

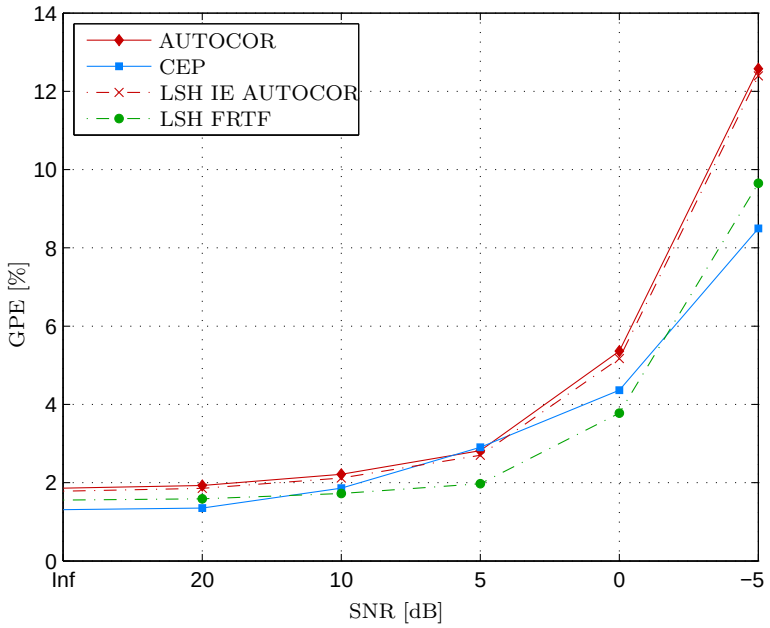
Conversely, the AUTOCOR method seems to be more suitable for female speech. Due to its dependency, the LSH IE AUTOCOR also performs quite bad, but is nevertheless slightly superior to its preceding estimator.

In the consideration without multiple pitch errors (Figure 4.5b), the LSH FRTF version shows the best performance. For clean speech signals, the GPE is identical to the one of the CEP method, but in noisy environments, it is distinctly smaller than the GPEs of the other methods.

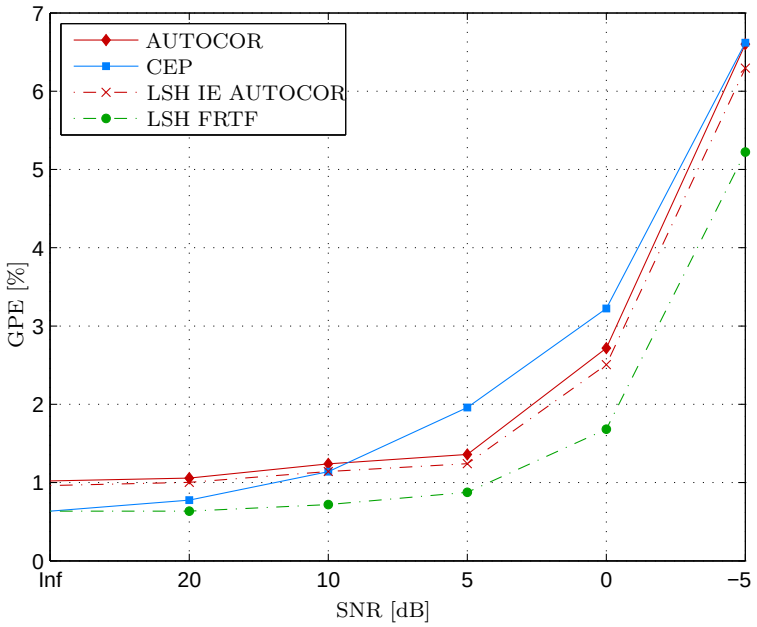
In Figure 4.6, the results for both genders are summarized. The initial estimate of the LSH IE AUTOCOR can, as already seen in the previous figures, always be refined, where the difference between the GPEs of the initial estimate and the LSH version is increasing with the noise level. Additionally, the GPE of the LSH FRTF version shows a relatively slow rise in comparison to the CEP method. Thus, indicating robustness of the LSH method against additive white Gaussian noise. For clean speech signals, there are only slight differences between the PDAs, expect the GPE of the LSH FRTF is noticeably higher. In noisy environments, the LSH IE AUTOCOR is the best performing-algorithm.

After excluding multiple pitch errors (Figure 4.6b), the GPE of the LSH FRTF version is closer to the ones of the AUTOCOR and LSH IE AUTOCOR methods, being competitive at very low SNR levels.

To sum up, the LSH pitch tracker seems to be robust against additive white Gaussian noise. A possible explanation for this is the MSE criterion, as mentioned in Section 3.3.2. Furthermore, the LSH IE AUTOCOR detector is suited to refine its initial estimate, where the improvement in accuracy becomes more pronounced with increasing noise level. However, the dependency on the initial estimate is a disadvantage whenever the preceding PDA is inaccurate, e.g. the AUTOCOR method in male speech. If multiple pitch errors can be eliminated, the independent version LSH FRTF is competitive at high noise levels.

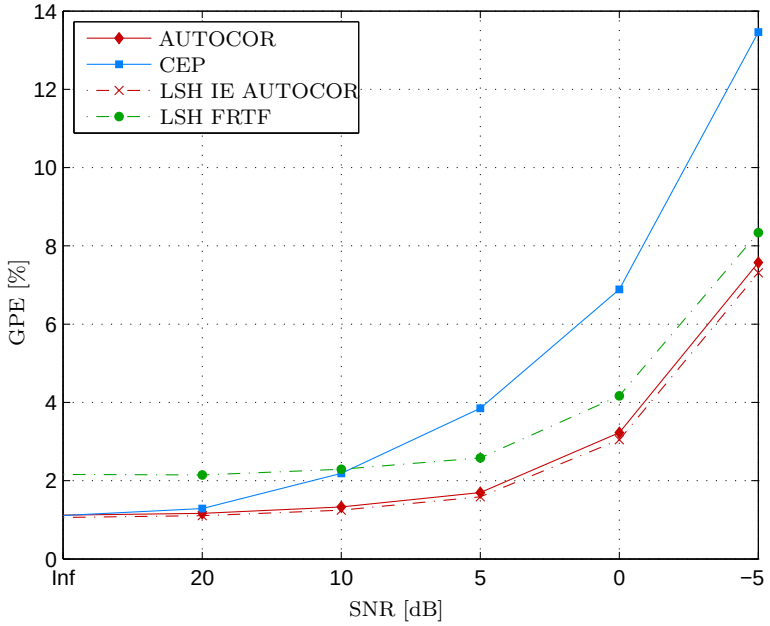


(a) Dependency of the average GPE on different SNR levels.

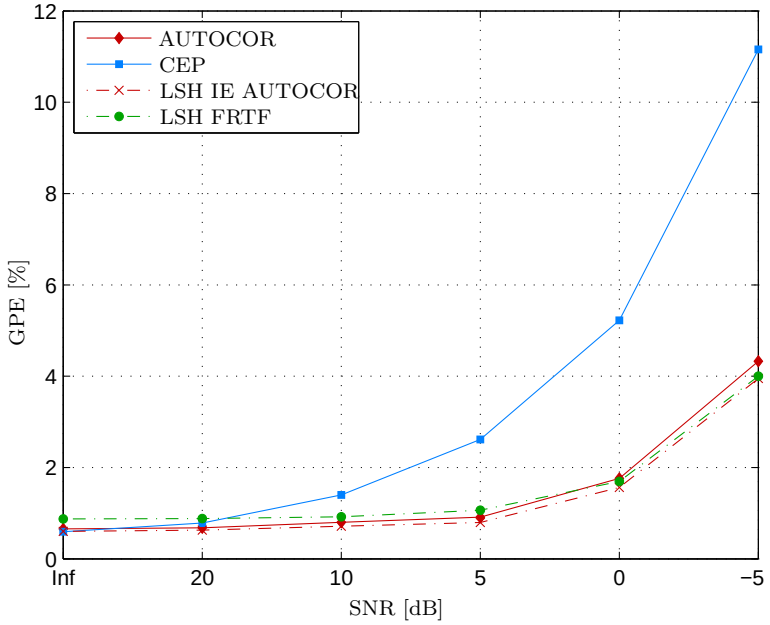


(b) Dependency of the average GPE on different SNR levels without frames containing multiple pitch errors.

Figure 4.5: Performance of the PDAs at different SNR levels for *male speakers*.



(a) Dependency of the average GPE on different SNR levels.



(b) Dependency of the average GPE on different SNR levels without frames containing multiple pitch errors.

Figure 4.6: Performance of the PDAs at different SNR levels for *all speakers*.

4.3.2 Robustness against Wind Noise

In the second scenario, wind noise recorded with an omnidirectional AKG CE20/17 microphone capsule that has been exposed to the airflow of a fan[2] is added to the Keele speech signals at $SNR = \{0, -10\}$ dB. Figure 4.7 shows a spectrogram segment of such wind noise bursts. From this figure, one can see the main characteristics of wind noise. Firstly, it is, in contrast to white Gaussian noise, of non-stationary nature [10] since the wind bursts occur randomly over time and secondly, the energy is concentrated in the lower frequencies.

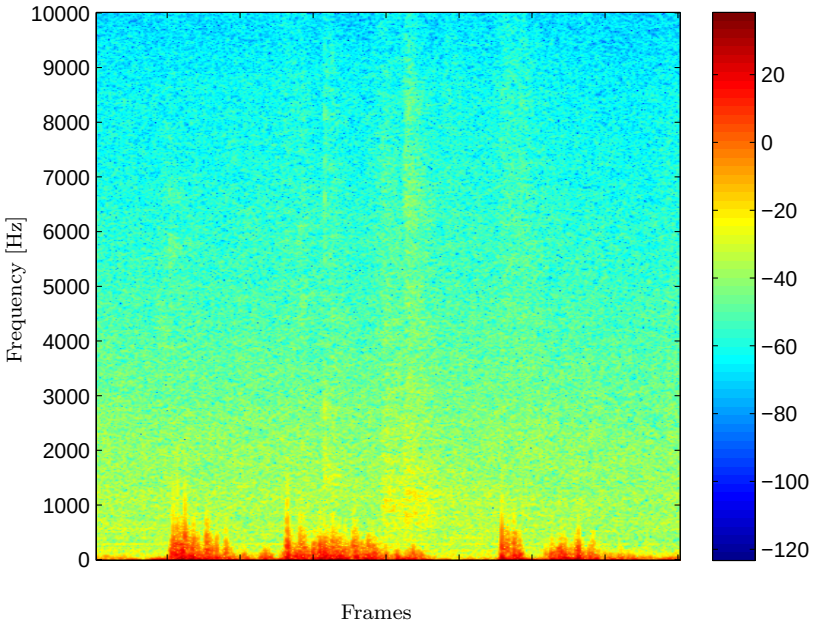


Figure 4.7: Spectrogram segment of wind noise bursts.

The results of the pitch trackers for this scenario are compared in the bar charts in Figures 4.8 to 4.10. It is apparent from these figures that the GPE increases with higher noise level for all algorithms. In the following, female and male speakers will also be discussed separately at first.

[2]This leads to turbulent airflow on the microphone membrane, which causes the characteristic wind noise.

For female speech (Figure 4.8), the LSH FRTF performs worst of all investigated algorithms. Although the LSH IE AUTOCOR is able to refine its initial estimate, it is far from being competitive to the CEP due to its strong dependency on the AUTOCOR method that shows a bad performance. In contrast to the white Gaussian noise scenario, the CEP method has a robust behavior in wind noise environments, as can be seen in the nearly constant GPE from 0 to -10 dB. This could not be exploited by using the CEP as initial estimate. Only a marginal refinement with the LSH IE CEP at 0 dB and no refinement at all at -10 dB was possible.

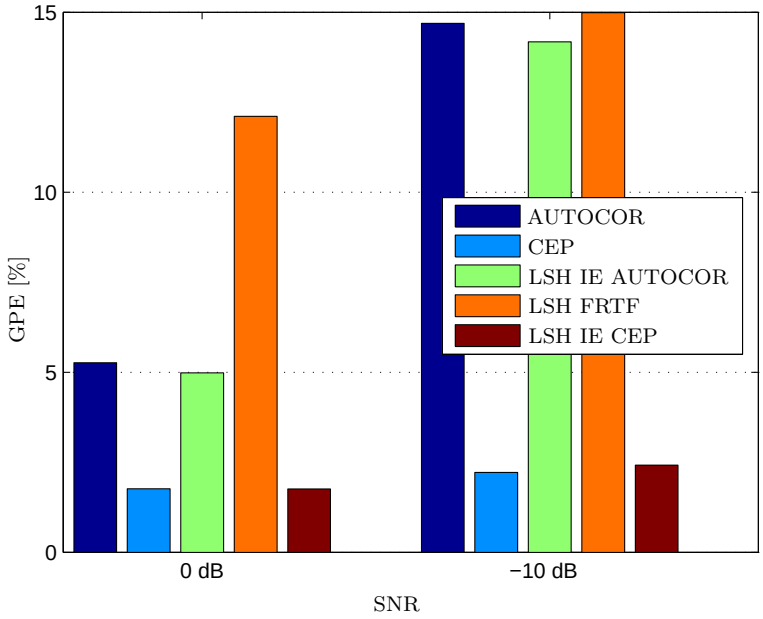


Figure 4.8: Performance of the PDAs in wind noise at different SNR levels for *female speakers*.

The estimation for male speakers, depicted in Figure 4.9, is more affected of the wind noise than for female speakers since more harmonics are located in the wind noise because of their usually lower pitch. As a consequence, the AUTOCOR method performs even worse, again leaving no chance for the LSH IE AUTOCOR. In comparison, the LSH FRTF seems to be distinctly more accurate in male speech, but is still not competitive to the CEP method. The CEP is not only the best-performing algorithm

for female speech, but also for male speech. Additionally, the LSH IE CEP cannot considerably reduce the GPE of its preceding estimator, too.

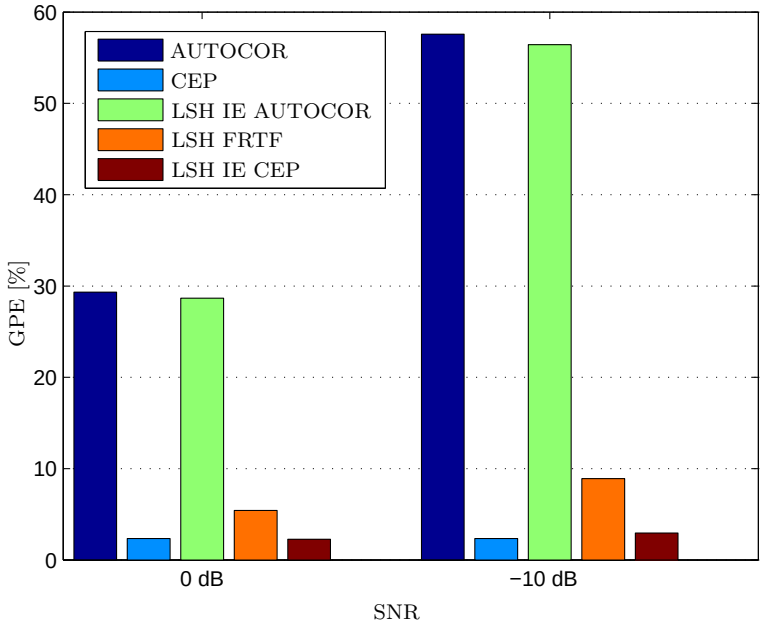


Figure 4.9: Performance of the PDAs in wind noise at different SNR levels for *male speakers*.

Figure 4.10 summarizes both results. The precise data for all bar charts discussed in this section can be found in Table A.4. It should be mentioned that the results of the observation without frames containing multiple pitch errors do not differ from the ones presented in this section with respect to the conclusions. The reason for this is that the differences between the GPEs of the algorithms are too high as to allow a change in the performance order by slight changes in the GPEs when excluding frames with multiple pitch errors.

The findings suggest that the LSH does not seem to be the best choice under wind noise conditions. The independent version LSH FRTF performs relatively bad and the estimation accuracy of the CEP pitch tracker cannot be improved in the LSH IE CEP version. An attempt to reduce the GPE of the LSH IE CEP by previous lowpass filtering of the noisy speech signals at 3 kHz, as done in [10], in order to eliminate the

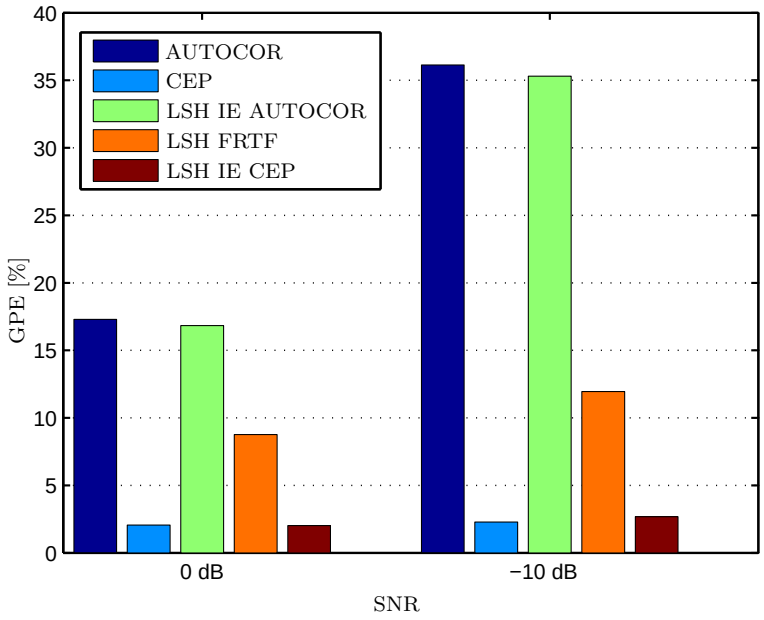


Figure 4.10: Performance of the PDAs in wind noise at different SNR levels for *all speakers*.

influence of unvoiced speech parts, which are concentrated in higher frequencies, was not very encouraging.

4.3.3 Computation Time

Another important factor in the comparison of the PDAs is their speed of execution. Since the algorithms were implemented in MATLAB, the recorded computation times are not meant to be interpreted as absolute values, but to provide a rough overview. Table 4.3 compares the computation time of the algorithms per second of speech sampled at 20 kHz. The LSH method is very costly due to its repetitive search over the range of possible pitch frequencies, as already mentioned in Section 3.3.2. Its short versions, the LSH IE and LSH FRTF, are four orders of magnitude slower than the two conventional methods. The LSH CIF version takes even another order of magnitude longer.

Table 4.3: Computation time per second of speech sampled at 20 kHz. The algorithms are implemented in MATLAB R2012b running on a single core of an Intel(R) Core(TM) i7 CPU 920 @2.67 GHz.

Pitch Detector	**Computation Time/s speech**
AUTOCOR	205 ms
CEP	147 ms
LSH IE AUTOCOR	39 s
LSH IE CEP	38 s
LSH FRTF	16 s
LSH CIF	5 min

Chapter 5

Conclusions & Future Work

The aim of this thesis was to evaluate the LSH algorithm, a Pitch Detection Algorithm that is claimed to show an accurate performance in noisy environments, which is still a desired feature in pitch estimation after many years of research. In the process, two versions were examined: the LSH Version Initial Estimate (LSH IE) and the LSH Version First Rough Then Fine (LSH FRTF). They were chosen due to their reduced computation time in comparison to the LSH Version Complete Interval Fine (LSH CIF) that did not result in any performance improvement despite the enormous additional computational cost. After preliminary tests on the dependency of the algorithm on its input parameters, e.g. the modeled number of harmonics, the versions were compared with two classical PDAs, the AUTOCOR and CEP methods, with respect to estimation accuracy, robustness against noise, and computation time. For this purpose, two scenarios were evaluated, white noise and wind noise conditions.

In the experiments the following strengths and weaknesses of the LSH algorithm have been worked out.

The *strengths* are:

- The method is a valid PDA providing very accurate pitch contours for clean

speech signals.

- It has indeed shown to be robust against additive white Gaussian noise. While the LSH IE version was able to refine its initial estimate, what is more noticeable at low SNR, the LSH FRTF version, outperformed in clean and mildly noisy environments, was starting to get competitive at very low SNR.

The *weaknesses* are:

- The algorithm does not seem most suitable for wind noise conditions. It was inferior to the CEP method, as the LSH IE version was not able to improve the accuracy at low SNR. The performance of the LSH FRTF was distinctly worse compared to the CEP method.

- Additionally, the LSH pitch detector is very costly in terms of computation time in comparison with the other methods. The improvements in accuracy in white noise environments must be interpreted carefully in relation to the enormous additional computation time.

- The limitation of the search interval and the resulting dependency on the initial estimate in the LSH IE version can also be a disadvantage whenever the preceding PDA is highly inaccurate.

- The whole-interval search (LSH FRTF version) involves a higher risk for estimation errors, like multiple pitch errors.

Additionally, it should be noted that there can be significant problems with numerical accuracy during solving the two sets of linear equations due to ill-conditioned matrices. This problem has been solved for the specified sampling rate and window length, but might occur again with other values of these parameters.

To sum up, the findings support the recommendation to use an initial estimate to limit the search interval and prevent potential errors. Other preceding estimators,

which need not be extremely accurate but do not make gross estimation errors, might further increase the performance of the LSH method. Moreover, the algorithm has shown encouraging results in white noise environments, but further work needs to be done to reduce the computation time.

A starting point could be to store the input signal-independent matrices $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{S}$ and $\mathbf{T}$ and the terms $cos[i\omega_0 k]$ and $sin[i\omega_0 k]$, required to calculate the input signal-dependent matrices $\mathbf{Y_1}$ and $\mathbf{Y_2}$, for each frequency between 60 Hz and 400 Hz with a step size of e.g. 0.1 Hz (see Section 3.3.2). Since these signal-independent steps altogether take about 80% of the total computation time. Note that they must be stored for a specific sampling rate, window size and number of harmonics.

Appendix A

Tables of Evaluation Results

The following tables provide the exact numbers of the charts discussed in Chapter 4. In Tables A.2 to A.4 the minimum GPE of the algorithms for each SNR level, i.e. the best performance, is highlighted in gray.

Table A.1: Dependency of the GPE for the LSH algorithm on the number of harmonics.

(a) Performance of the LSH algorithm for *all speakers*.

GPE in %	P = 5	P = 10	P = 15	P = 20
LSH IE AUTOCOR	1.18	1.07	1.06	1.06
LSH FRTF	2.39	1.92	2.16	2.60

(b) Performance of the LSH algorithm for *female speakers*.

GPE in %	P = 5	P = 10	P = 15	P = 20
LSH IE AUTOCOR	0.35	0.34	0.34	0.34
LSH FRTF	0.90	2.31	2.76	3.41

(c) Performance of the LSH algorithm for *male speakers*.

GPE in %	P = 5	P = 10	P = 15	P = 20
LSH IE AUTOCOR	2.01	1.80	1.79	1.79
LSH FRTF	3.87	1.52	1.56	1.79

Table A.2: Robustness against additive white Gaussian noise in terms of the GPE.

(a) Performance of the PDAs at different SNR levels for *all speakers*.

GPE in %	AUTOCOR	CEP	LSH IE AUTOCOR	LSH FRTF
Inf dB	1,12	1,11	1,06	2,16
20 dB	1,16	1,29	1,11	2,14
10 dB	1,33	2,19	1,25	2,29
5 dB	1,70	3,85	1,59	2,58
0 dB	3,23	6,89	3,04	4,17
-5 dB	7,57	13,46	7,31	8,34

(b) Performance of the PDAs at different SNR levels for *female speakers*.

GPE in %	AUTOCOR	CEP	LSH IE AUTOCOR	LSH FRTF
Inf dB	0,38	0,91	0,34	2,76
20 dB	0,40	1,22	0,36	2,70
10 dB	0,44	2,52	0,38	2,86
5 dB	0,57	4,79	0,48	3,19
0 dB	1,09	9,41	0,92	4,56
-5 dB	2,57	18,43	2,22	7,03

(c) Performance of the PDAs at different SNR levels for *male speakers*.

GPE in %	AUTOCOR	CEP	LSH IE AUTOCOR	LSH FRTF
Inf dB	1,86	1,31	1,78	1,56
20 dB	1,93	1,35	1,85	1,59
10 dB	2,21	1,86	2,12	1,73
5 dB	2,82	2,90	2,70	1,97
0 dB	5,36	4,36	5,17	3,78
-5 dB	12,58	8,50	12,40	9,65

Table A.3: Robustness against additive white Gaussian noise in terms of the GPE with exclusion of frames containing multiple pitch errors.

(a) Performance of the PDAs at different SNR levels for *all speakers*.

GPE in %	AUTOCOR	CEP	LSH IE AUTOCOR	LSH FRTF
Inf dB	0,66	0,60	0,60	0,87
20 dB	0,68	0,78	0,63	0,88
10 dB	0,80	1,40	0,71	0,92
5 dB	0,91	2,61	0,80	1,06
0 dB	1,76	5,22	1,56	1,70
-5 dB	4,33	11,16	3,95	4,00

(b) Performance of the PDAs at different SNR levels for *female speakers*.

GPE in %	AUTOCOR	CEP	LSH IE AUTOCOR	LSH FRTF
Inf dB	0,29	0,56	0,24	1,11
20 dB	0,30	0,79	0,25	1,13
10 dB	0,36	1,66	0,29	1,12
5 dB	0,46	3,27	0,35	1,25
0 dB	0,80	7,23	0,62	1,71
-5 dB	2,06	15,70	1,61	2,78

(c) Performance of the PDAs at different SNR levels for *male speakers*.

GPE in %	AUTOCOR	CEP	LSH IE AUTOCOR	LSH FRTF
Inf dB	1,02	0,63	0,96	0,63
20 dB	1,06	0,77	1,00	0,64
10 dB	1,24	1,14	1,14	0,72
5 dB	1,36	1,96	1,24	0,88
0 dB	2,72	3,22	2,51	1,68
-5 dB	6,60	6,62	6,29	5,22

Table A.4: Robustness against wind noise in terms of the GPE.

(a) Performance of the PDAs at different SNR levels for *all speakers*.

GPE in %	AUTOCOR	CEP	LSH IE AUTOCOR	LSH IE CEP	LSH FRTF
0 dB	17,29	2,05	16,83	2,02	8,77
-10 dB	36,14	2,28	35,31	2,68	11,95

(b) Performance of the PDAs at different SNR levels for *female speakers*.

GPE in %	AUTOCOR	CEP	LSH IE AUTOCOR	LSH IE CEP	LSH FRTF
0 dB	5,26	1,77	4,98	1,76	12,11
-10 dB	14,69	2,22	14,18	2,42	14,99

(c) Performance of the PDAs at different SNR levels for *male speakers*.

GPE in %	AUTOCOR	CEP	LSH IE AUTOCOR	LSH IE CEP	LSH FRTF
0 dB	29,32	2,34	28,67	2,27	5,42
-10 dB	57,59	2,34	56,43	2,95	8,90

Appendix B

UML Diagram Pitch Detection Algorithms

Figure B.1: UML Diagramm Pitch Detection Algorithms.

Appendix C

Abbreviations and Acronyms

ACF	Autocorrelation Function
AUTOCOR	Autocorrelation Method
CEP	Cepstrum Method
GPE	Weighted Gross Pitch Error
H+N	Harmonic-Plus-Noise
LSH	Least Squares Harmonic
LSH CIF	LSH Version Complete Interval Fine
LSH FRTF	LSH Version First Rough Then Fine
LSH IE	LSH Version Initial Estimate
MATLAB	MATrix LABoratory (C) The Mathworks, Inc.
MMSE	Minimum Mean Squared Error
MPER	Multiple Pitch Error Rate
MSE	Mean Squared Error
PDA	Pitch Detection Algorithm
SNR	Signal-to-Noise Ratio
UML	Unified Modeling Language

Appendix D

Notation

D.1 Notation in General

f_s	sampling frequency
k	sample index
ω	normalized frequency
f_0	fundamental (or pitch) frequency
τ_0	pitch period
N_{FFT}	FFT length
P	number of harmonics in the H+N model

D.2 Mathematical Operators

$DFT\{\cdot\}$	Discrete Fourier Transform
$IDFT\{\cdot\}$	Inverse Discrete Fourier Transform
$FFT\{\cdot\}$	Fast Fourier Transform

Bibliography

[1] N. Abu-Shikhah and M. Deriche, "A robust technique for harmonic analysis of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, May 2001, pp. 877–880.

[2] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical v/uv classification algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 333–338, May 1999.

[3] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," in *Third European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 2, 1993, pp. 1003–1006.

[4] I. Bronstein, K. Semendjajew, G. Musiol, and H. Muehlig, *Taschenbuch der Mathematik*. Frankfurt am Main: Verlag Harri Deutsch GmbH, 2008.

[5] F. Flego, "Pitch Estimation," http://www.icocla.it/keele.html, [Online, accessed 02-August-2013].

[6] C. Hofmann, "Suppression of impulse-like noises in a microphone signal," Diploma thesis, University of Erlangen-Nuremberg, 2011.

[7] T. Holt, "Stephen Hawking," http://www.hawking.org.uk/about-stephen.html, 2013, [Online, accessed 03-September-2013].

[8] W. Kellermann, "Signal Processing for Speech and Audio," Lecture notes, University of Erlangen-Nuremberg, Summer term 2012.

[9] W. Kellermann, "Digital Signal Processing," Lecture notes, University of Erlangen-Nuremberg, Winter term 2012/2013.

[10] B. King, "Enhancing single-channel speech in wind noise using coherent modulation comb filtering," Master's thesis, University of Washington, 2008.

[11] D. A. Krubsack and R. J. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 319–329, Feb. 1991.

[12] MathWorks, "Condition Number," http://www.mathworks.de/de/help/matlab/ref/cond.html, 2013, [Online, accessed 06-August-2013].

[13] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[14] R. J. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Apr. 1990, pp. 249–252.

[15] A. M. Noll, "Cepstrum pitch determination," *The Journal of the Acoustical Society of America (JASA)*, vol. 41, no. 2, pp. 293–309, 1967.

[16] A. M. Noll, "Clipstrum pitch determination," *The Journal of the Acoustical Society of America (JASA)*, vol. 44, no. 6, pp. 1585–1591, 1968.

[17] F. Plante, Meyer G., and W. Ainsworth, "A pitch extraction referene database," in *Fourth European Conference on Speech Communication and Technology (EUROSPEECH)*, Sept. 1995, pp. 837–840.

[18] Qin Li and L. Atlas, "Time-variant least squares harmonic modeling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Apr. 2003, pp. II – 41– 4.

[19] T. F. Quatieri, *Principles of discrete-time speech processing.* Upper Saddle River, NJ, USA: Prentice Hall, 2001.

[20] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 1, Feb. 1977.

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

[22] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals.* Englewood Cliffs, NJ, USA: Prentice-Hall, 1978.

[23] B. Secrest and G. R. Doddington, "Postprocessing techniques for voice pitch trackers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 7, May 1982, pp. 172–175.

[24] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 727–730, Oct. 2001.

[25] Y. Stylianou, "Modeling speech based on harmonic plus noise models," in *Nonlinear speech modeling and applications*, G. Chollet, Ed. Berlin and New York: Springer, 2005, vol. 3445, pp. 244–260.

[26] P. Vary, U. Heute, and W. Hess, *Digitale Sprachsignalverarbeitung.* Stuttgart: Teubner Verlag, 1998.

[27] V. Viswanathan and W. Russell, "New objective measures for the evaluation of pitch extractors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 10, Apr. 1985, pp. 411–414.