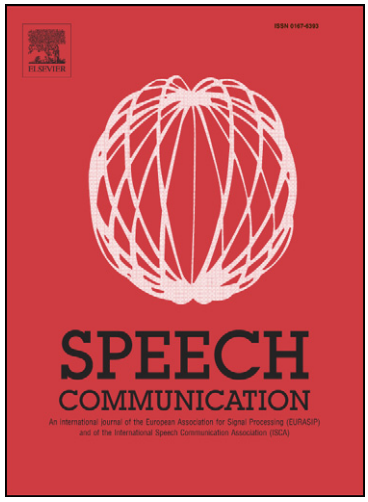
On the Detection of Pitch Marks Using a Robust Multi-Phase Algorithm

M. Legát, J. Matoušek, D. Tihelka



On the Detection of Pitch Marks Using a Robust Multi-Phase Algorithm

M. Legát, J. Matoušek, D. Tihelka

PII: S0167-6393(11)00009-4
DOI: 10.1016/j.specom.2011.01.008
Reference: SPECOM 1966

To appear in: *Speech Communication*

Received Date: 9 October 2009
Revised Date: 19 August 2010
Accepted Date: 14 January 2011

Please cite this article as: Legát, M., Matoušek, J., Tihelka, D., On the Detection of Pitch Marks Using a Robust Multi-Phase Algorithm, *Speech Communication* (2011), doi: 10.1016/j.specom.2011.01.008

On the Detection of Pitch Marks Using a Robust Multi-Phase Algorithm

M. Legát, J. Matoušek*, D. Tihelka

Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Univerzitní 8, 306 14 Pilsen, Czech Republic

Abstract

A large number of methods for identifying glottal closure instants (GCIs) in voiced speech have been proposed in recent years. In this paper, we propose to take advantage of both glottal and speech signals in order to increase the accuracy of detection of GCIs. All aspects of this particular issue, from determining speech polarity to handling a delay between glottal and corresponding speech signal, are addressed. A robust multi-phase algorithm (MPA), which combines different methods applied on both signals in a unique way, is presented. Within the process, a special attention is paid to determination of speech waveform polarity, as it was found to be considerably influencing the performance of the detection algorithms. Another feature of the proposed method is that every detected GCI is given a confidence score, which allows to locate potentially inaccurate GCI subsequences. The performance of the proposed algorithm was tested and compared with other freely available GCI detection algorithms. The MPA algorithm was found to be more robust in terms of detection accuracy over various sets of sentences, languages and phone classes. Finally, some pitfalls of the GCI detection are discussed.

Key words: glottal closure instant, pitch mark, speech signal polarity, fundamental frequency

*Corresponding author

Email addresses: legatm@kky.zcu.cz (M. Legát), jmatouse@kky.zcu.cz (J. Matoušek)

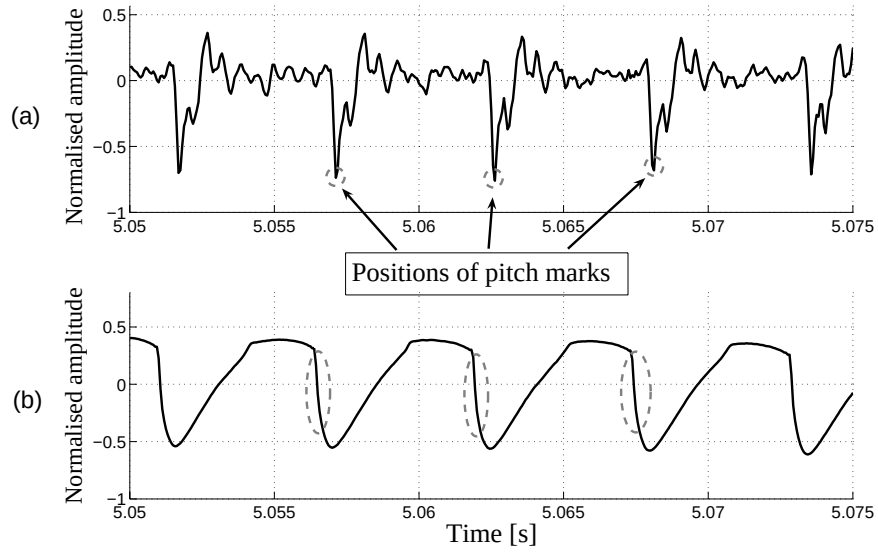


Figure 1: (a) short segment of voiced speech signal, (b) corresponding EGG waveform with highlighted glottal closures.

1. Introduction

The modern pitch-synchronous methods of speech processing [1] rely on the knowledge of moments of the glottal closures. These moments are called glottal closure instants (GCIs) or pitch marks. Pitch mark can be defined as the location of a speech signal amplitude extreme (peak or valley) that corresponds to the moment of glottal closure. In Fig. 1 part (a), a segment of voiced speech is shown, where the positions of pitch marks are depicted and part (b) demonstrates the corresponding electroglottograph (EGG) signal, in which positions of glottal closures are highlighted. The electroglottograph is a device which enables non-invasive measurement of the time variation of the contact between vocal cords without affecting speech production. For more detailed information refer to [2]. It is obvious that pitch marks are present only in voiced segments of speech as there is no vocal fold vibration in unvoiced segments of an utterance.

Pitch marks are generally used in pitch-synchronous speech synthesis methods (e.g. PSOLA, some kinds of sinusoidal synthesis, etc.) [3]-[6], where they ensure that speech is synthesized in a consistent manner. The con-

catenation of waveforms somewhere in between prominent amplitude events, corresponding to fundamental frequency periods, is likely to result in perceived phase discontinuities, which is a problem often mentioned in OLA-based methods [7]. Also, no matter that units are concatenated in pitch-synchronous way, inconsistent labeling (one with pitch marks assigned to peaks and the second to valleys, or vice versa) leads to phase discontinuities.

Pitch marks can also be utilized in a number of speech analysis and processing methods including the calculation of cycle-based analysis measures [8], pitch-synchronous speech enhancement [9], automatic phonetic segmentation [10] or design of concatenation cost functions in unit selection speech synthesis [11]. Another application is in clinical diagnosis and treatment of voice pathologies. Knowing positions of glottal closures, a very accurate estimation of an f_0 contour can be obtained and used in voice conversion techniques [12], [13] or for intonation recognition [14].

However, the detection of glottal closure instants in speech has two major drawbacks: it is not very robust if performed automatically, and time-consuming if performed manually [15]. An alternative possibility is to measure the glottal activity directly using an electroglottograph (also called laryngograph or EGG), as mentioned above, and derive positions of pitch marks from its recordings.

It should be noted at this point that the current freely available pitch marking algorithms are inclined to be error-prone even if glottal signals are available, which is the case of many speech corpora intended for the purposes of high-quality speech synthesis.

Besides EGG and speech signals, LPC residual signals, despite suffering from some generally known imperfections, can also be used for pitch marking, especially in cases when EGG recordings are not available. Let us remark, however, that using the glottal signals, superior results can be expected since they are not burdened by modifications that happen to a flow of speech in the vocal tract.

In recent years, various methods of pitch marking have been proposed, including the wavelet-based analysis [16]–[18], the application of nonlinear system theory [19], threshold-based and/or peak picking methods [20]–[23], and group delay methods [24]–[27]. This list of related works could be extended, but it is not limited to [28]–[31].

Before any pitch marking algorithm is employed, it needs to be decided whether the pitch marks should be placed at peaks or at valleys of a speech waveform. As discovered during our experiments, this decision is very impor-

tant for the performance of a pitch marking algorithm in terms of its accuracy and robustness. In [32] the problem of peak/valley decision making is solved by comparing the f_0 contour calculated using AMDF (Average Magnitude Difference Function) and f_0 contours derived from valley and peak-based pitch mark sequences. The decision depends on the deviation between these contours. This particular issue was also addressed in [33] or [34].

In Section 2, we propose a simple method for peak/valley decision making based on a confrontation of peaks and valleys of a speech waveform [35].

The next step is to find accurate locations of pitch marks. As mentioned above, extracting pitch marks from the EGG signals is much simpler than using speech waveforms alone. In Section 3, we briefly describe a common method to detect pitch marks, henceforth referred to as "the baseline algorithm" (BLA), that was formerly used at our department, and which is highly dependent on a quality of the EGG signals. Note that these signals tend to suffer from imperfections caused, for example, by improper placement of measuring electrodes on a speaker's throat due to an inconvenient shape of his/her larynx. Some imperfections may also occur during the production of voiced fricatives when the pressure ratios in the vocal tract cause imperfect closing of vocal cords, which is normally not observable in the speech signal itself. The same phenomenon can be observed in aspirated speech segments (see Fig. 6). Thus, the detection of pitch marks based on the EGG signals is not always as straightforward as it may seem. We have observed some considerable weaknesses of BLA resulting from such EGG signals imperfections, which led us to the implementation of a new method described in this paper.

In Section 4, steps of a new robust algorithm based on several methods of pitch tracking and pitch marking [36] are described. This algorithm works in several phases, utilizes a dynamic programming routine to find the best sequence of pitch marks and uses both the EGG and the corresponding speech signals, taking advantage of feasible features of both. The novelty of this algorithm lies within a unique combination of known approaches resulting in a robust method characterized by high accuracy of the pitch mark detection across various speakers, languages and recording conditions.

Section 5 serves to demonstrate the performance of the proposed method in comparison to other freely available algorithms. In addition, the issue of variance of delays between EGG and corresponding speech signals is also addressed in that section.

Finally, in Section 6 we draw some conclusions and outline our plans for future work.

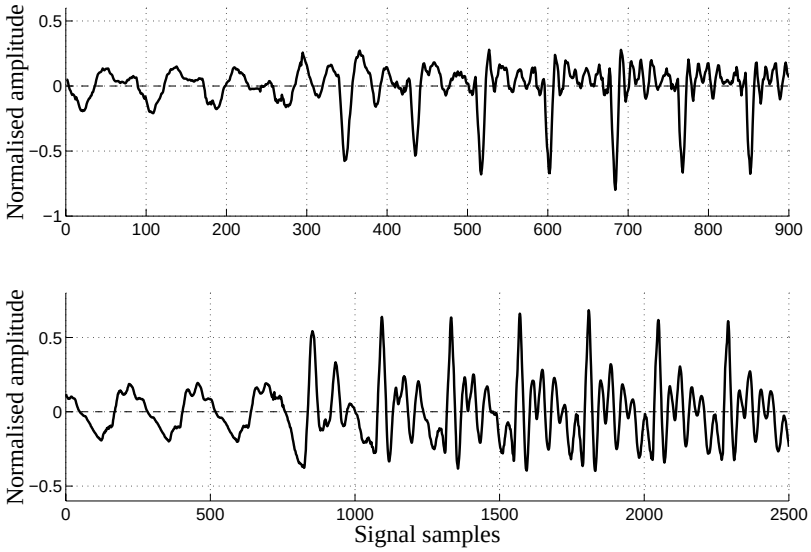


Figure 2: Polarity mismatch. In the upper part there is a segment of the Sentence1 (negative polarity), while in the lower part there is a segment cut from the Sentence2 (positive polarity).

2. POLARITY OF SPEECH SIGNAL

2.1. Motivation

During the development of our pitch marking algorithm, a large variation was observed in its performance. We have discovered that this was to a great extent due to a polarity mismatch present in our speech corpus [37]. This mismatch is illustrated in Fig. 2, where speech segments extracted from two sentences of different dominant polarity are shown. This observation led to an idea to develop the peak/valley decision making method that is described in the following subsection.

The reason why the performance of any pitch marking algorithm can be affected by setting the wrong polarity is that there can be more than one dominant signal amplitude extreme in the vicinity of a prospective location of a pitch mark, which results in a "spurious" jitter. To make matters worse, the local extreme which corresponds to the moment of glottal closure does not need to be the one with the largest amplitude. The simplest way of avoiding such an ambiguity is to determine the dominant polarity of a speech signal before pitch marking itself because then there is, in most cases, only one or

two at most dominant signal amplitudes to choose from when placing pitch marks.

2.2. Peak/Valley Decision Making Method

As an input to the polarity detection algorithm, the maximum f_0 of the speaker needs to be estimated. Since the recordings contained in a speech corpus for the concatenative speech synthesis are made by a single speaker, this task can easily be accomplished manually. The estimate does not need to be very accurate, because, as described below, it only serves to set the signal in a vicinity of selected peaks to zero.

The proposed method can be summarized as follows. Firstly, the speech waveform needs to be pre-processed. The aim of the pre-processing is to reduce higher frequencies present in unvoiced segments and any extraneous noise. This is accomplished by the low-pass filtering of the input speech signal. The goal of this filtering is to remove high frequencies and preserve the valleys and peaks in voiced segments (see Fig. 3), which is necessary for later stages of the algorithm.

Having the pre-processed speech waveform, the next step of the proposed method is to confront speech signal peaks and valleys. In this confrontation we use both the pre-processed speech waveform (*speech*) and its absolute value (*abs_speech*):

$$abs\_speech = |speech| \,. \tag{1}$$

The method can be summarized as follows:

1. Initialize counters *peak_count* and *valley_count* with zero values.
2. Find the global maximum of the *abs_speech*. Denote its time coordinate as t_m.
3. If the position of this maximum corresponds to a position of a peak in the *speech*, increment the counter *peak_count*, otherwise the *valley_count* is incremented.
4. To remove other peaks in the vicinity of t_m, set the value of *abs_speech* to zero in the range:

$$[t_m - 2/3 * T_0, t_m + 2/3 * T_0], \tag{2}$$

where $T_0 = 1/f_0$ and f_0 is the estimate of a speaker's maximum value of the fundamental frequency acquired manually. The signal is set to zero within the given range to avoid selection of spurious signal amplitude

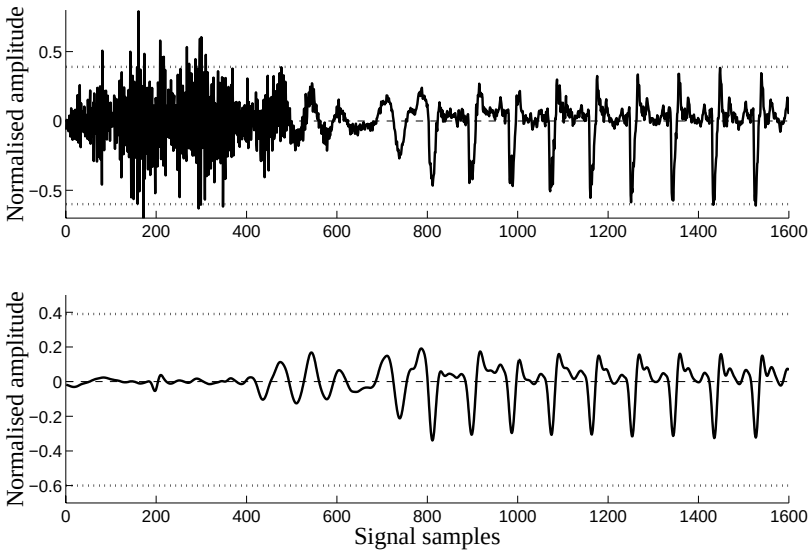


Figure 3: Raw and pre-processed speech waveform. The dotted lines serve to illustrate how noisy speech segments could influence the peak/valley decision. The signal amplitude extremes in noisy frames, which exceed the upper line or decrease below the lower line, would be erroneously counted, which would considerably affect the performance of the algorithm.

extremes, which do not correspond to glottal closures, and to allow for selection of neighbouring "true" peaks/valleys in the next iteration.

5. Repeat steps 2, 3 and 4 until the RMS energy value of the *abs_speech* is lower than $thresh * rms\_speech$, where the *rms_speech* is the RMS energy value of the *speech*. The recommended value of the *thresh* constant lies in the range $[0.2, 0.7]$, the higher this value is the faster the peak/valley decision is made.

In fact, the algorithm only compares signal peaks and valleys in terms of their amplitudes. For the final peak/valley decision, we also calculate the overall energy above *e_above* and below zero *e_below* of the signal *speech*. These energy values are used as auxiliary predictors. If the *peak_count* is higher than the *valley_count* and *e_above* is higher than *e_below*, peaks are believed to be convenient for pitch marks placement and vice versa. In the case that the values of the counters are not in accordance with the values of energies, only the counters are used to make a decision. Such decisions are

Table 1: Summary of experiment results. The values in the table are accuracies of automatic pitch marking in percentage. "Peak" means peak-based pitch marks, "Valley" means valley-based pitch marks. The polarity of tested sentences was negative.

	Peak	**Valley**
CZ-M2	88.18	98.10
CZ-F	87.20	97.74
SK-F	88.21	97.19
GE-M	86.04	91.01

then marked as uncertain.

2.3. Evaluation of Peak/Valley Decision Making Method

Rather than experiments, some results of a practical utilization of the proposed method are presented in the first part of this subsection. The method was employed to check and unify the polarity of the newly recorded speech corpus [37]. The corpus was built specially for the purposes of unit-selection text-to-speech synthesis as it consists of a large number of both phonetically and prosodically rich sentences, their recordings (both speech and glottal signals) and both orthographic and phonetic annotations (totally, 12,277 utterances were recorded, almost 18hours of speech excluding pauses). More details about the corpus can be found in [37]. The results were more than satisfactory — 98.14% correct decisions, 1.36% correct but uncertain decisions and only 0.5% errors.

In addition, an experiment was designed to measure how the peak/valley decision influences a performance of a pitch marking algorithm. For this purpose we used the MPA algorithm described in Section 4; nevertheless, the other methods listed in Section 5.2 behave similarly regarding wrong polarity selection for pitch marking. The accuracy of the algorithm was tested with respect to the polarity of pitch mark positions — either peaks (local maxima) or valleys (local minima) of speech waveforms. The experiment was conducted in three languages — Czech (CZ-M2 male and CZ-F female), Slovak (SK-F) and German (GE-M). In 8 sentences, 2 of each set, the pitch marks were placed manually by the authors (see Sec. 5.1 for more details) resulting in a set containing about 7,000 reference pitch marks. The sentences were randomly selected from corpora recorded for the purposes of concatenative speech synthesis. In each sentence the pitch marks were placed at peaks and valleys separately. Hence, we obtained two pitch mark sequences

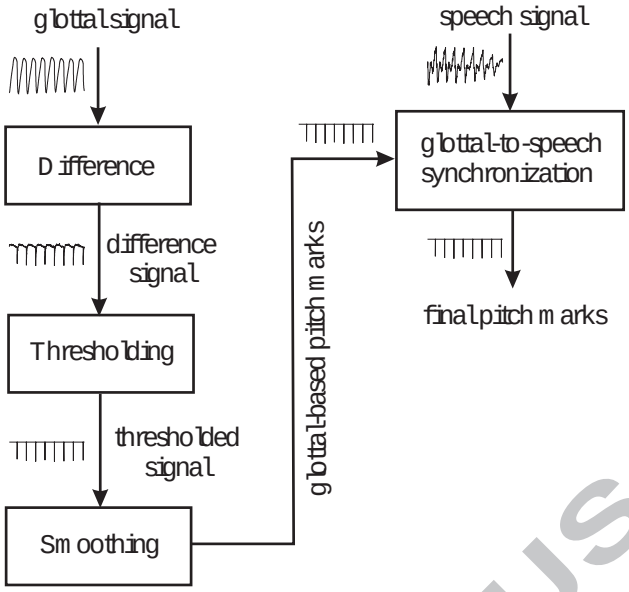


Figure 4: Block diagram of the baseline pitch mark detection algorithm.

— peak-based pitch marks and valley-based — for each sentence. Note that the polarity of all sentences used in this experiment was the same and it was negative (i.e. valleys are more appropriate for pitch marks placement).

The summary of the results can be seen in Tab. 1. The average loss of accuracy if the pitch marks were placed with incorrect polarity setting (i.e. placing to peaks when the overall polarity of speech signals was negative) was 8.6%. The accuracy was measured using the equation (7), which is explained in Section 5.3.

3. BASELINE ALGORITHM AND TASKS TO TACKLE

In the first part of this section we describe a fast and simple pitch marking algorithm based on the thresholding of the difference function applied to the glottal signals, henceforth referred to as Baseline Algorithm (BLA). In the second part, we outline its main weaknesses, which were the motivation for the development of a new algorithm, described in the next section.

3.1. Description of the Baseline Algorithm

As glottal (EGG) signal (capturing the vocal fold vibration during speaking) is very suitable for the purposes of GCI detection, it is often used as

the input of the pitch marking systems. Despite its useful properties, it is desirable to highlight its features with some preprocessing. In our system, the difference of the glottal signal is used as the input of the pitch marking algorithm. The difference signal emphasizes the moments of glottal closures since it reflects the rate of changes of the status of the vocal cords. There are usually large sharp negative peaks in the difference signal which correspond to the moments of glottal closures. Moreover, using the difference signal, phenomena caused by improper vertical placing of the electrodes across the neck (not centered on the larynx) are also reduced.

The core of the baseline algorithm is in a thresholding of the difference signal which, in fact, removes slow changes of the status of the vocal cords. All positive values and values lower than a given threshold (the global RMS value of the difference signal in our case) are set to zero. All nonzero values in the thresholded signal form then a set of candidates for the pitch marks placement.

The next procedure performs pitch mark smoothing. It examines the pitch mark candidates in the thresholded signal and tries to avoid doubling and halving pitch mark placements by removing the pitch mark candidates within one pitch period (i.e. the candidates which are too close to each other), "isolated" candidates (surrounded by regions of unvoiced speech) and possibly also by inserting pitch marks which were removed incorrectly by previous thresholding. After the pitch marks smoothing stage, the nonzero samples that remain in the "smoothed" signal correspond to the positions of GCI in the glottal signal.

The last step of the baseline algorithm takes into account a time shift between glottal and speech signal (see Fig. 7), which appears due to the difference in positions where these signals are acquired. While the glottal signal is measured using electrodes placed around the speaker's larynx, the speech signal is acquired by a microphone positioned in front of his/her lips. The delay tends to vary with speakers and depends on how far the speaker is from the microphone and on the velocity of the air flow going through the vocal tract, i.e. on the speaker's vocal tract dimensions.

To obtain the positions of pitch marks, which is crucial for concatenative speech synthesis systems, the GCIs detected in glottal signal need to be shifted into speech. This shift needs to be consistent across all utterances to avoid phase mismatches at concatenation points at synthesis runtime, and that is why signal peaks or valleys are reasonable choices. Many methods consider this shift between glottal and speech signal static but the opposite

is the case. In the previous versions of our pitch mark detection system, an average static shift of $500\mu s$ was performed to cope with the time lag but it resulted in phase mismatches at concatenation points. In the BLA, the performance of which is presented in this paper, a range for searching for an extreme in the speech signal was set as $\langle t_s, t_s + 0.3T_0 \rangle$, where T_0 is the local estimate of a fundamental pitch period and t_s is the time of the detected GCI. More details about the baseline pitch mark detection algorithm, excluding dynamic shift, can be found in [38]. The scheme of the BLA is shown in Fig. 4.

3.2. Tasks to tackle

As the BLA proved to be very sensitive to a quality of the glottal signals, a more robust algorithm had to be developed. We illustrate the main weaknesses of the BLA in the following paragraphs.

3.2.1. False Peaks in the Course of Difference Function

Fig. 5 shows a section of the glottal signal accompanied by the corresponding thresholded difference function. The false peaks in the course of the thresholded difference function and the corresponding edges in the course of the glottal signal are highlighted by solid boxes. Unfortunately, these false peaks cannot be removed by thresholding because it would lead to the removal of correct peaks in the segments of low amplitude glottal signal as well. This failure causes placement of spurious pitch marks.

3.2.2. Imperfect Closing of Vocal Cords

Another problem of the method based on difference function is demonstrated in Fig. 6. It is quite obvious that pitch marks should be placed in all speech signal valleys in the shown voiced section. Examining the glottal signal and the corresponding thresholded difference function, we can see that no pitch mark candidates are generated in the second half of the section shown in the figure. This course of EGG waveform is due to the imperfect closing of vocal cords, as mentioned in Section 1.

3.2.3. Shift between Glottal and Speech Signal

As stated above, there is a time shift between glottal and speech signals varying with speakers (to be precise, their vocal tract dimension), and the placement of the measuring equipment. Some additional shift may be introduced by mastering (a process of speech signal enhancement based on

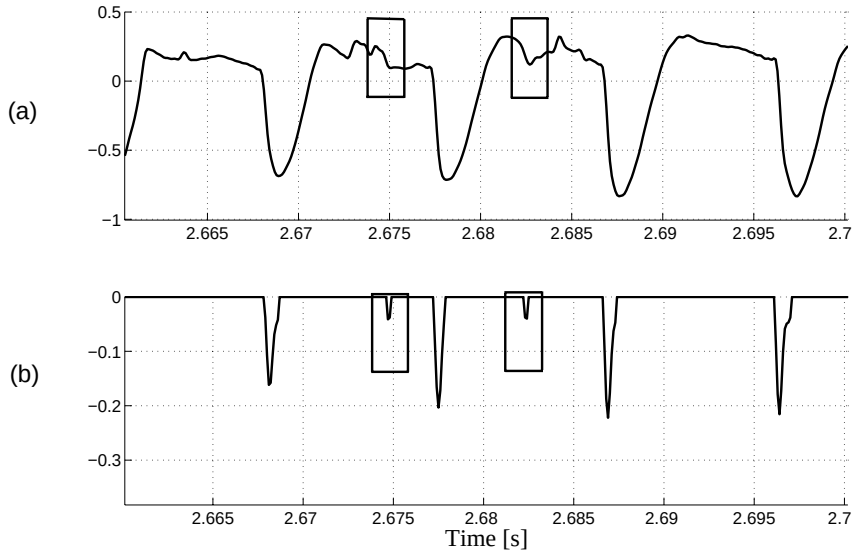


Figure 5: (a) EGG waveform, (b) Thresholded difference of EGG waveform. Solid boxes serve to highlight false peaks in thresholded difference function which are caused by imperfection of EGG signal.

filtering) of recorded signals. The typical value of the delay is less than 1 ms, according to our experiments, but sometimes much longer delays may be observed (see the right part of Fig. 7 and Tab. 6), which the BLA is not able to handle.

4. A ROBUST MULTI-PHASE PITCH MARK DETECTION ALGORITHM

In this section we describe the design of a new algorithm for pitch marking — a robust multi-phase pitch mark detection algorithm (MPA). This algorithm is based on a unique combination of several known approaches, resulting in a more robust and consistent method, which seems to outperform other available pitch marking algorithms (see Sec. 5.4). The main idea is to utilize both glottal and speech signal for pitch marking. The scheme of the MPA is shown in Fig. 8.

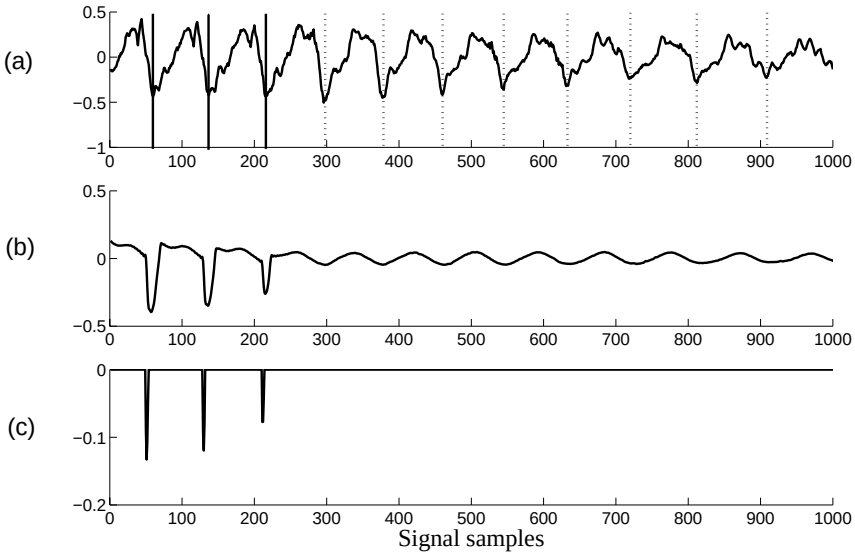


Figure 6: (a) Speech waveform with pitch marks, (b) EGG waveform, (c) Thresholded difference of EGG waveform. In part (a), the missing pitch marks are highlighted by dotted lines. In part (b), the imperfect closing of vocal cords is demonstrated in EGG waveform (aspirated speech in this case).

4.1. Voicing estimation

The very first step of the proposed method is the estimation of voicing. The voicing contour is very important for the placement of pitch marks in segments where no edges are found in the glottal waveform (see Fig. 6). In such segments, the speech waveform is used to find a sequence of prospective pitch mark candidates. We make use of two voicing contours in further steps of the algorithm. One is obtained directly from the glottal waveform and one from the filtered speech waveform.

The two voicing contours are used as a base to define a confidence measure. There is a high confidence in pitch marks placement in locations which are voiced according to both of these contours. If any frame is marked as voiced according to only one of the voicing contours, the confidence is lower. The least, but still some, confidence is obtained for frames which are voiced only according to the EGG based contour. It is obvious that there is no confidence in placing pitch marks in frames which are marked as unvoiced in both contours. The advantage of having two voicing contours obtained

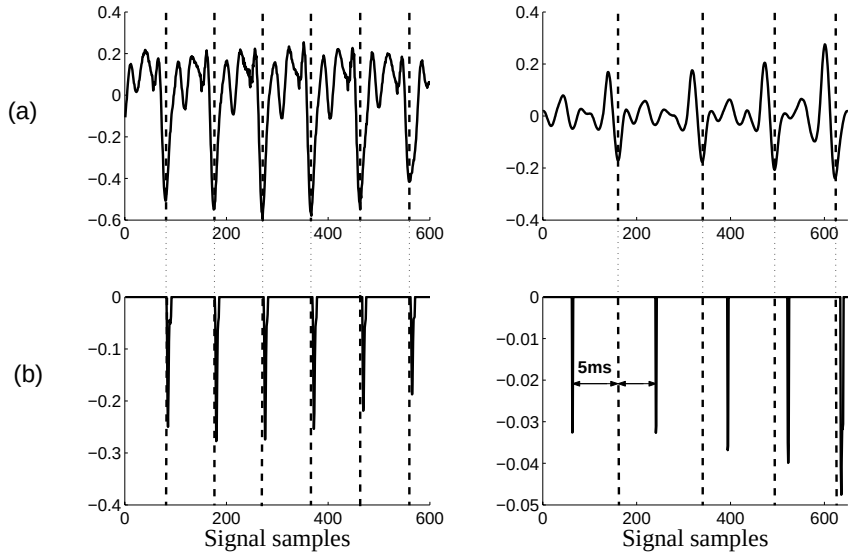


Figure 7: (a) Speech waveform with pitch marks, (b) Thresholded difference of EGG waveform. Delays between speech and glottal waveform can vary in a large range. In the left part there is almost no delay, while in the right part the time lag is about 5 ms.

in different ways is also that the voicing estimation method can fail in some frames and thus it is more robust to use both EGG and speech waveform for estimation, but separately.

For the EGG based frame voicing classification we use values of short-time energy function, zero-crossing function and YIN algorithm [39] as predictors. To obtain a voicing contour based on the speech waveform, the YIN algorithm is used solely. Both voicing contours are evaluated on short-time basis, the window length is 20 ms and the overlap is a half of the window length. One of the useful features of the YIN method is that each estimate is given a confidence and we take advantage of it in both voicing and f_0 estimation, which is the next step of our algorithm.

4.2. Fundamental frequency estimation

Having obtained the voicing contour, the next step is to estimate the f_0 contour. We have decided to utilize the YIN algorithm [39] for this purpose, as one of its features is that the f_0 estimate for each frame is given a confidence, so that uncertain estimates can be subsequently removed from

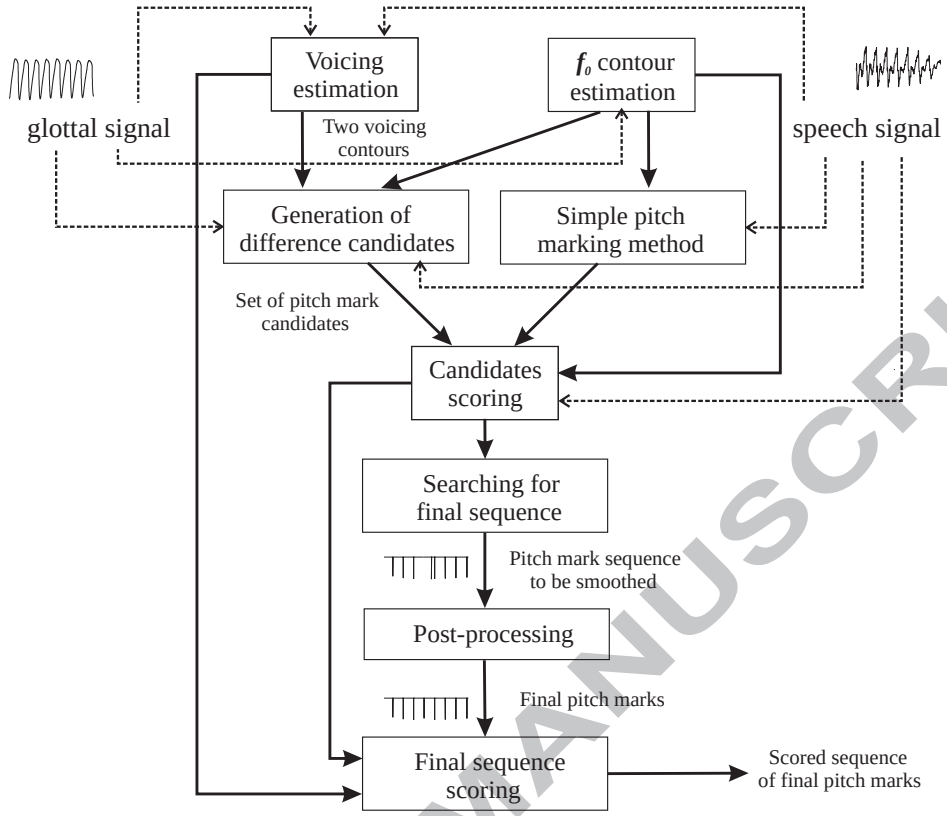


Figure 8: Block diagram of the multi-phase pitch mark detection algorithm.

the contour. The YIN algorithm is based on the well-known autocorrelation method which is modified in order to enhance its performance. In our implementation, a glottal waveform is used as the input, because it does not contain higher frequencies.

In the frames where the YIN algorithm gives uncertain results the thresholded difference function of glottal signal is utilized. The f_0 estimate is found as an inverse of the median value of differences between peaks of the difference function. This estimation is used only if the number of peaks in a given frame is higher than a minimum number $n_{\min}$ (this value depends on the frame length and the expected f_0 range — for $f_0 \approx 200$ Hz we use $n_{\min} = 4$).

Sometimes both of these approaches fail even if the given frame is voiced. In this case, the f_0 estimate is found by interpolation using f_0 values in the neighbouring frames.

4.3. Generation of pitch mark candidates

Having the estimates of both pitch and voicing contours, we continue with the generation of pitch mark candidates. In this phase, two approaches are applied to obtain a set of pitch marks candidates containing three candidate positions for each prospective pitch mark.

4.3.1. Modified difference function method

We use the modified BLA to generate a sequence of "difference" candidates. The modification consists in pitch mark smoothing (see Sec. 3.1) when the pitch contour estimated in the previous phase is taken into account instead of applying general thresholds for detection of "doubled" and "isolated" candidates. This allows us to partially solve the problems of false peaks described in Sec. 3.2.1.

Since there is also the problem of missing candidates (see Sec. 3.2.2), the voicing contours and the low-pass-filtered speech signal are used to add missing candidates into the "difference" sequence. The objective of filtering is to avoid generating candidates in unvoiced speech segments. The procedure of generating additional candidates can be summarized as follows:

1. The first and the last candidate is found in every subsequence of "difference" candidates. These subsequences are separated by unvoiced intervals.
2. To find the prospective successor of the last candidate in the given subsequence, the search region defined as

$$[c^{(i)} + 0.8T_0^{(i)}, c^{(i)} + 1.4T_0^{(i)}], \tag{3}$$

   is explored, where $c^{(i)}$ is the last candidate in a subsequence and $T_0^{(i)}$ is the local estimate of the pitch period.
3. If there is a peak in this region, the amplitude of which is higher than the voiced threshold thr_{v}, it is added into the candidate sequence. This procedure is repeated until all peaks to the right of the last candidate are found. In the same way, all predecessors of the first candidate in the subsequence can be found.

The constants 0.8 and 1.4 (step 2) were found by a grid search mechanism maximizing the overall accuracy (7) on the half of the testing set presented in this paper and their incorporation was motivated by a need to handle minor imperfections in the local pitch period estimates and also jitter. In Fig. 9,

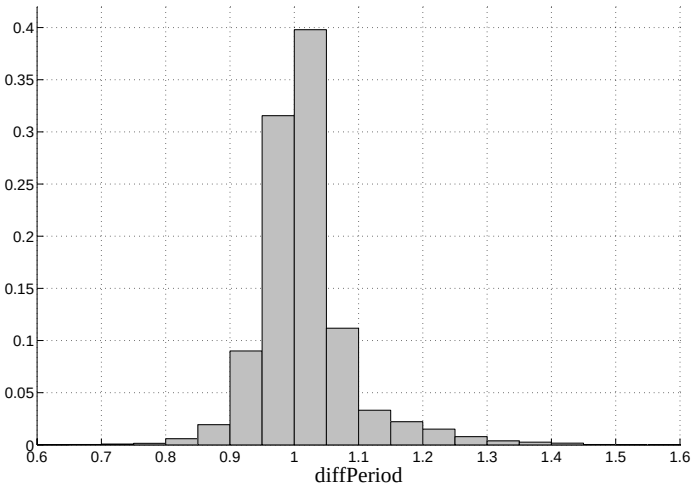


Figure 9: Distribution of the ratios of lengths of consecutive pitch periods. The *diffPeriod* was defined as *diffPeriod*=p_{i+1}/p_i, where p_i is the length of the i-th pitch period.

the distribution of ratios of all consecutive pitch periods extracted from the manually labeled data is shown, explaining partly why these particular values were found to be the best.

The value of constant thr_{v} depends on the voicing state of the frame in which the candidate is to be placed, as described in Sec. 4.1. Based on the manually placed pitch marks we found the optimal range for this threshold as $thr_{\mathrm{v}} \in \langle 0.35 * RMS, 1.1 * RMS \rangle$.

4.3.2. Simple pitch marking method

To generate two more candidates for each final pitch mark, we use the Simple Pitch Marking Method (SPM) [32]. In the case that there are two local extremes of the speech signal in a region where one final pitch mark is to be placed, and the lower one reaches at least 80% of the amplitude of the higher one, both of them are used as candidates. Otherwise, only the highest one is used as a doubled candidate.

The advantage of this simple method is that all prospective pitch mark positions are marked. If we put "difference" candidates and candidates obtained by SPM together, we have a set of candidates in which the final pitch mark sequence can be found. Note that we have at least three candidate positions for each final pitch mark (either three single, or one doubled and

one single, or, very rarely, one tripled).

4.4. Finding of the final pitch mark sequence

Before we start searching for the final sequence, all candidates need to be given a score. The scoring of candidates is based on two characteristics of pitch marks: the amplitude of speech signal at pitch mark position should be large (amplitude score) and the sequence of pitch marks should correspond to the estimated f_0 contour (position score).

We define a relative amplitude measure of each candidate i in a search region j as:

$$\bar{s}_i(j) = \frac{h_i(j)}{h_{\max}}, \tag{4}$$

where $h_i(j)$ is the amplitude of the candidate and $h_{\max}$ is the absolute value of the local extreme in the search region. This measure is then normalized [32] so that the sum of amplitude scores of candidates in a given search region is equal to 1.

The position score of two candidates is defined in the same way as described in [32]

$$\bar{t}(i,j) = \frac{1}{1 + |f_0 - \frac{f_s}{d}|}, \tag{5}$$

where f_0 is the local estimate of fundamental frequency; f_s is the sampling rate and d is the distance between candidates in samples. Using this measure the position score of the candidate j_1 in search region i and candidate j_2 in search region $i+1$ can be defined [32], as follows :

$$t_i(j_1, j_2) = \frac{\bar{t}_i(j_1, j_2)}{\sum_{k=1}^{3} \bar{t}_i(j_1, k)}. \tag{6}$$

The total candidate score is then defined as the sum of its amplitude score, position score and the amplitude score of its successor. It means that each candidate in the search region has three scores as it has three prospective successors. The amplitude scores of the successors are included because of the presence of "doubled" candidates. In the case that two candidates are placed at the same position, their amplitude scores are the same. Their final score would be then affected only by position scores towards their successors,

which is sensitive to values of f_0 taken from the overall contour. Including the amplitude scores of successors leads to higher robustness of the final score. Once the sequence of candidates is scored, the final sequence with the highest total score is searched for by means of dynamic programming, similarly as in [32].

4.5. Post-processing

Occasionally, some errors in the final pitch mark sequence may occur. This is because we use the sequence of "difference" candidates to determine search regions for final pitch marks, which may result in "doubled" and "missing" errors.

Fortunately, the detection of these errors is very simple. We use two thresholds to find these errors in pitch mark sequence, "doubled" threshold $thr_{\mathrm{d}} = 0.5$ and "missing" threshold $thr_{\mathrm{m}} = 1.5$. If the distance between two successive pitch marks is higher than $thr_{\mathrm{m}} \cdot f_0$ or less than $thr_{\mathrm{d}} \cdot f_0$, the missing or doubled error is detected respectively.

4.6. Scoring of the final sequence

The advantage of the proposed algorithm is that, along with the final pitch mark sequence, a score sequence is generated as well. Consequently, we have some information about how the final pitch mark sequence was generated and which pitch marks are likely to be uncertain.

The final pitch mark score is derived from the amplitude candidate score. We raise this value by 5% if the final pitch mark comes from the original "difference" candidates sequence (i.e. sequence found only on the base of the glottal waveform). If the final pitch mark is placed in the voiced speech segment according to the speech based voicing contour (see Fig. 8), we raise its confidence by 10%. These raised percentage values were set experimentally.

To address the second demand on the pitch mark position (i.e. position with regards to its neighbours), we use the position candidate score. We separately score the distance between the final pitch mark and its predecessor and the distance between the final pitch mark and its successor.

5. EXPERIMENTS

In this section we present the results of experiments which were conducted to test the performance of the proposed method. The aim of these

experiments was to find a quantitative assessment of the performance of the proposed algorithm in comparison with some other, freely available, methods.

We have also done an experiment to find optimal time shifts for transitions between GCIs and selected amplitude extremes (peaks or valleys) in speech signal waveforms. The results are presented in paragraph 5.4.4.

5.1. Test material

To find out how the proposed algorithm performs across various languages, recording conditions and pitch ranges, we have conducted an experiment using nine sets of sentences spoken by various speakers — Czech female (CZ-F), Czech male (CZ-M1, CZ-M2), German male (GE-M), English male (EN-M), French female (FR-F), and Slovak female (SK-F). In addition, a set of sentences which contained "tasks to tackle" described in Section 3.2 (CZ-TT), and a set of sentences uttered in a happy speaking style (HA-F) were tested. The set of happy speaking style sentences was included because the dynamics of their f_0 contours is higher when compared to neutral sentences, which may make the pitch marking task more difficult. Each of the sets contained ten sentences, except CZ-TT, which contained only three sentences. Note that the sentences used for the evaluation of the peak/valley decision making method described in Sec. Sec:PeakValleyEval were taken from these sets.

All these sentences were labeled manually with pitch marks being a reference for our experiments. The total number of manually placed pitch marks was 51,200 in 83 sentences containing natural distribution of all phoneme categories. To ensure that the manually placed pitch mark sequences were unbiased, the pitch marks in three of those sentences were placed independently by two experts. The average discordance was only 4 pitch marks per sentence (including operations shift, insert and delete — see Sec. 5.3).

5.2. Methods tested

In our experiments we have compared the proposed algorithm (MPA), the baseline algorithm (BLA) described in Section 3.1, the program `pitchmark` implemented in Edinburgh Speech Tools Library[1] (EST), Speech Filing System[2] (SFS), and Praat [23].

[1] http://www.cstr.ed.ac.uk/projects/speech_tools
[2] http://www.phon.ucl.ac.uk/resource/sfs

Table 2: Description of the Praat methods used for pitch marking.

Input signal	Sound to Pitch	Pitch to PointProcess	Label
EGG	To Pitch...	To PointProcess (peaks)	PIP-G
Speech	To Pitch...	To PointProcess (peaks)	PIP-S
Speech	To Pitch...	To PointProcess (cc)	PIC-S
Speech	To Pitch (cc)...	To PointProcess (peaks)	PCP-S

SFS was tested in three ways. First, we found the instants of glottal closures in EGG waveform and moved these instants into positions of speech signal extremes in the same way as in BLA (SFS-G). Another set of pitch marks was found only on the base of knowledge of speech waveform and f_0 contour derived from glottal waveform (SFS-S). The last but not least SFS set of pitch marks (SFS-HQ) was determined using the `HQTx program`, the output of which was aligned with speech.

In the Praat system there are various methods to create a "Pitch Object" from a selected "Sound Object". Although the method "Sound: To Pitch..." is preferred for speech signals, we have also tested the others. A "Pitch Object" can then be converted into "PointProcess" to find locations of high amplitude using methods "Sound&Pitch: To PointProcess (cc)" or "Sound&Pitch: To PointProcess (peaks)". The latter method gives more variable periods [23]. In this paper we present only the results of the best four combinations (see Table 2). The performance of other methods was found significantly worse in our experiments.

5.3. Performance measure

To compare a sequence of manually detected pitch marks (S_R) with a sequence of automatically detected pitch marks (S_T), a dynamic-programming algorithm (modified Levenshtein distance of sequences of time instants) was employed. This algorithm searches for the minimum number of transformations needed to derive the sequence S_R from the sequence S_T. The transformations considered are substitution (S), deletion (D) and insertion (I) [38]. The accuracy of automatic pitch mark detection is defined as follows:

$$Accuracy = \frac{N_R - N_S - N_D - N_I}{N_R} \times 100[\%], \tag{7}$$

where N_R is the number of pitch marks in the reference sequence S_R, N_S is the number of substitutions, N_D is the number of deletions and N_I is

Table 3: Summary of the experiment results on Czech language. "Acc" means accuracy and the values of N_D, N_I and N_S are in cases per sentence.

CZ-M1	**Acc**[%]	N_D	N_I	N_S	**CZ-F**	**Acc**[%]	N_D	N_I	N_S
EST	85.41	54.4	3.6	9.1	EST	84.07	97.2	7.4	13.7
BLA	94.78	1.5	17.0	4.4	BLA	94.76	1.4	25.9	6.4
SFS-G	**96.28**	7.0	7.9	2.6	SFS-G	90.59	8.8	22.7	14.6
SFS-S	77.57	106.7	13.5	4.4	SFS-S	87.07	99.7	18.7	8
SFS-HQ	91.16	36.2	1.4	5.3	SFS-HQ	83.37	60.0	3.8	20.95
PIP-G	89.81	38.8	1.7	9.3	PIP-G	89.86	68.7	2.3	11.9
PIP-S	89.39	33.8	10.0	3.8	PIP-S	90.59	22.6	13.7	9.8
PIC-S	88.74	23.3	18.7	7.4	PIC-S	87.28	26.8	21.6	28.0
PCP-S	88.67	35.0	9.7	5.7	PCP-S	91.34	43.5	13.4	9.4
MPA	96.19	5.4	5.0	4.6	MPA	**97.39**	7.9	5.1	5.7
CZ-M2	**Acc**[%]	N_D	N_I	N_S	**CZ-TT**	**Acc**[%]	N_D	N_I	N_S
EST	88.03	87.6	5.5	11.5	EST	73.25	79.7	20.7	35.7
BLA	94.48	3.93	19.2	7.7	BLA	93.28	0.7	13.4	8
SFS-G	97.97	7.1	16.7	11.3	SFS-G	54.45	8.3	32	177.7
SFS-S	80.81	97.5	19.3	7.0	SFS-S	69.15	49.7	33.3	74.7
SFS-HQ	94.37	51.4	2.8	14.5	SFS-HQ	54.67	70.3	14.7	150
PIP-G	94.14	54.9	2.9	10.4	PIP-G	91.90	29	1	19.3
PIP-S	92.67	39.7	13.4	8.9	PIP-S	90.59	34	3.7	12
PIC-S	90.39	22.3	20.9	29.2	PIC-S	58.06	19.7	9.7	165
PCP-S	93.37	38.6	11.7	8.3	PCP-S	93.01	26	3	7
MPA	**98.02**	7.6	4.0	4.8	MPA	**98.42**	2.3	2.7	5

the number of insertions involved in the comparison process. If a distance between a pitch mark in S_R and S_T is lower than 10% of a local pitch period, no penalty is given. The value 10% seems to be reasonable since the results presented in [1] suggest that such a pitch marks misplacement does not affect the quality of synthetic speech (speaking about PSOLA-based synthesis).

5.4. Experiment Results

5.4.1. Overall accuracy comparison

The results of our experiments performed on neutral sentences across various languages and speakers are summarized in Tab. 3 and Tab. 4. In Tab. 5, the results obtained on sentences uttered in a happy speaking style are shown. For clarity, we also present a box and whisker plot, see Fig. 10,

showing the comparison of the overall performance of all the tested methods across all sentences. Since the notch of the MPA plot does not overlap with any other, there is a strong evidence [40] that the MPA outperforms other algorithms under evaluation in terms of the overall accuracy (7).

It is also obvious from the tables that the MPA algorithm is more accurate than other methods on all sets except the CZ-M1 and the FR-F, where the SFS-G and the BLA methods, respectively, give slightly better results. Nevertheless, the difference in performance between these two methods and the MPA algorithm is not statistically significant on these sets. The results suggest that the MPA algorithm is very robust as its lowest performance was 93.25% on French data.

Regarding the set of sentences uttered in a happy speaking style, we had expected somewhat inferior performance of all the methods due to the higher dynamics of the f_0 contours but this hypothesis was supported by the obtained results only in part. Surprisingly, the performance of methods PIP-G, PIP-S and PCP-S was slightly better than on neutral sentences.

The obtained accuracies of methods SFS-G and SFS-HQ are also notable as the performance of these methods was significantly deteriorated by the nature of the sentences. Nevertheless, we have found that their performance can be improved by setting a larger range for searching amplitude extremes in the glottal-to-speech synchronization phase.

Regarding the gender of a speaker, MPA tends to give similar results in terms of accuracy even if the EGG signals of male and female speakers are different. However, no definite conclusion can be drawn as our results are language-biased.

5.4.2. Accuracy across various phoneme categories

In order to see how the proposed method works across various phone categories, we defined seven phone classes for testing — *vowels*, *diphthongs*, *plosives*, *fricatives*, *affricates*, *nasals* and *others*. All the tested methods were inclined to follow the same accuracy trend where the *nasals* was the class with the highest obtained accuracies reaching 99.5%, whereas the *affricates* and the *plosives* were found to be the most difficult classes for labeling with pitch marks. The worst performance of the MPA algorithm was observed on *affricates* falling to 72%. Note that the performance of the other methods was even worse on this set, which might be explained by the noisiness and the instability of these sounds.

Table 4: Summary of the experiment results obtained for German, English, French and Slovak. "Acc" means accuracy and the values of N_D, N_I and N_S are in cases per sentence.

GE-M	**Acc**[%]	N_D	N_I	N_S	**FR-F**	**Acc**[%]	N_D	N_I	N_S
EST	85.91	76.6	4.5	11.1	EST	87.89	77.6	6.2	15.5
BLA	86.97	3.7	22.4	6.2	BLA	**93.83**	3.8	19.7	13.9
SFS-G	92.26	7.2	14.4	11.1	SFS-G	88.63	12.6	14.9	18.0
SFS-S	83.97	83.9	16.8	6.1	SFS-S	73.74	78.1	18.8	27.6
SFS-HQ	93.23	43.9	2.2	11.7	SFS-HQ	89.59	42.4	4.0	14.8
PIP-G	89.77	45.0	4.2	10.4	PIP-G	91.78	43.6	4.3	13.6
PIP-S	86.74	36.4	13.7	8.0	PIP-S	91.69	33.8	13.8	12.0
PIC-S	87.49	20.7	21.1	23.4	PIC-S	65.70	20.7	21.6	56.8
PCP-S	85.41	37.1	11.6	7.4	PCP-S	92.64	38.7	11.9	9.6
MPA	**94.25**	8.2	4.1	4.2	MPA	93.25	8.7	3.8	11.6
EN-M	**Acc**[%]	N_D	N_I	N_S	**SK-F**	**Acc**[%]	N_D	N_I	N_S
EST	70.90	82.2	3.8	10.1	EST	93.50	66.6	8.5	14.6
BLA	89.76	4.1	19.5	9.7	BLA	95.72	3.0	21.4	12.4
SFS-G	86.87	8.0	13.1	12.2	SFS-G	94.34	13.5	20.3	46.0
SFS-S	80.09	77.5	15.6	8.4	SFS-S	83.92	73.9	20.7	29.3
SFS-HQ	89.21	41.5	1.9	11.0	SFS-HQ	95.48	63.7	34.1	23.8
PIP-G	86.64	44.6	3.8	10.0	PIP-G	90.06	44.1	4.3	14.6
PIP-S	85.72	35.9	12.2	7.6	PIP-S	91.62	34.7	12.3	12.5
PIC-S	82.31	21.1	19.0	22.9	PIC-S	78.45	22.2	19.4	71.8
PCP-S	80.93	40.6	10.2	7.5	PCP-S	92.23	37.2	11.0	10.6
MPA	**94.73**	8.6	3.7	4.4	MPA	**97.30**	8.8	4.5	9.5

5.4.3. Accuracy vs. pitch range evaluation

Since the pitch range of the input signals is an important factor in the evaluation of any pitch marking algorithm, we present the performance of the four best methods with respect to the pitch range of the speakers, see Fig. 11. The pitch range was defined as follows:

$$f0Range = 4 * std\left(allF0es\right), \tag{8}$$

where $allF0es$ is a vector of all local $f0$ estimates for a given speaker calculated from three consecutive manually placed pitch marks. It is obvious that the performance of the MPA is very stable.

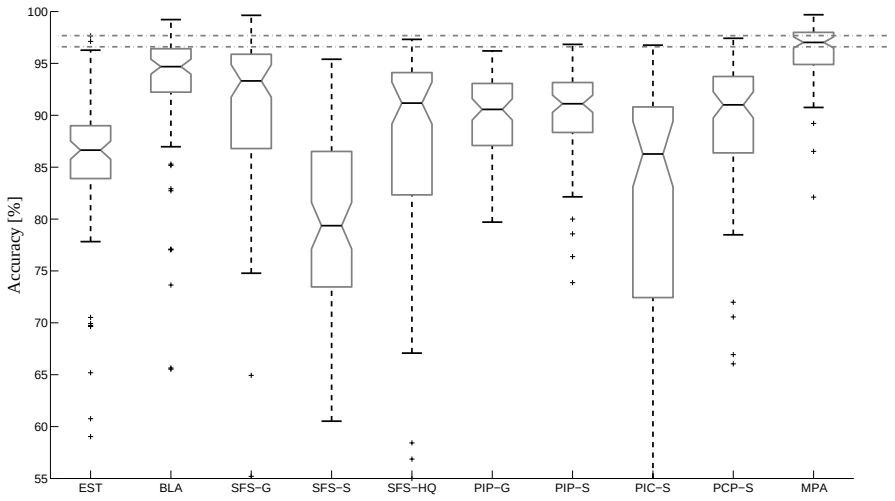


Figure 10: Overall performance of all the tested algorithms. The displayed values are the accuracies obtained from all the tested sentences, including the happy set. For most of the methods, these sentences were displayed as outliers since their performance was considerably worse on this data.

Table 5: Summary of the experiment results on sentences uttered in a happy speaking style. "Acc" means accuracy and the values of N_D, N_I and N_S are in cases per sentence.

HA-F	**Acc**[%]	N_D	N_I	N_S
EST	84.81	69.4	8.8	11.6
BLA	91.48	3.3	19.3	14.0
SFS-G	30.4	14.1	17.7	51.6
SFS-S	80.74	73.3	19.3	25.6
SFS-HQ	12.08	56.8	24.3	70.3
PIP-G	92.04	40.3	3.9	14.2
PIP-S	92.91	30.9	12.9	12.1
PIC-S	75.00	18.7	20.2	60.1
PCP-S	93.25	35.0	11.3	10.0
MPA	**97.26**	8.0	3.8	10.6

5.4.4. Time shifts between EGG and speech

For the experiments with time shifts between EGG and speech waveforms we have used the SFS-G method. We have found a sequence of GCIs for each

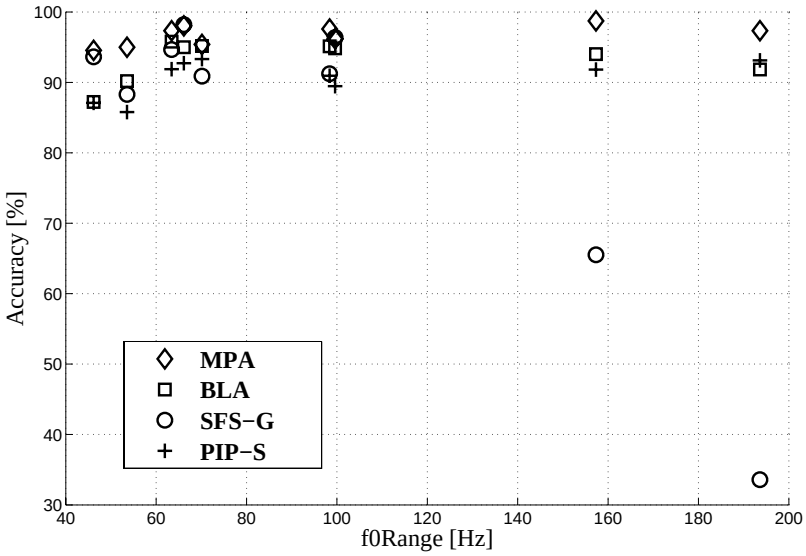


Figure 11: Overall performance of the four best methods with respect to the $f0$ range of the speaker. Note that the largest $f0$ range was observed for the HA-F set.

sentence and searched for the optimal range of time shift for the transition from GCIs to amplitudes of a speech waveform.

The results, summarized in Tab. 6, suggest that a delay between EGG and speech signals depends on the speaker and, obviously, on the placement of the measuring equipment. A surprising result has been obtained for the set CZ-M1, where the time shift was about three times longer than in other sets. It is also worth pointing out the comparison between results obtained on sets CZ-F and HA-F. These two sets contain utterances spoken by the same speaker, the only difference is the presence of portrayed happy emotion in the latter set.

6. CONCLUSIONS & FUTURE WORK

We have presented a robust multi-phase algorithm (MPA) which takes advantage of feasible features of both glottal and speech signals for pitch marking. The novelty of the proposed method consists in a unique combination of known methods resulting in a very robust and tunable algorithm, as well as in the utilization of two knowledge sources for pitch marking.

Table 6: Average optimal time shifts between EGG and speech

Set	**Mean shift**[ms]	**Std**
CZ-M1	2.39	0.48
CZ-M2	0.81	0.56
CZ–F	**0.69**	**0.41**
DE-M	0.72	0.72
EN-M	0.75	0.49
FR-F	1.39	0.45
SK-F	0.99	0.46
HA-F	**1.78**	**0.66**

Its performance was tested by experiments and compared with other freely available pitch marking algorithms. According to our experiments, the MPA algorithm seems to be more robust in terms of pitch marks placement accuracy over various sets of sentences, languages and phone classes than other methods. We have also conducted an experiment on a set of sentences uttered in a happy speaking style characterized by high f_0 range and dynamics. In contrast to the SFS-G method, which performed well on other sets and failed on these sentences, the performance of MPA was still rather stable. As the proposed MPA algorithm seems to be robust enough, it has been incorporated into our speech synthesis system [22].

Since the initial step of the pitch marking should be a speech polarity decision due to its importance for the performance of any pitch marking algorithm, we have also addressed this particular issue and proposed a simple method of peak/valley decision making.

The future work will focus on the optimization of the parameters involved in the MPA algorithm. The results presented in this paper were obtained under the general default parameter setting. All parameters, however, may be tuned automatically using the manually pitch marked data. The idea is to have task oriented sets of parameters, which would allow accurate pitch marking across various languages and also emotional states of speakers.

Acknowledgments

This research was supported by the Ministry of Education of the Czech Republic, project No. 2C06020, and by the Grant Agency of the Czech Republic, project No. GACR 102/09/0989.

References

[1] E. Moulines and F. Charpentier, *Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones*, Speech Communication, vol. 9(5–6), Dec. 1990, pp. 453–467.

[2] M. Rothenberg, *A multichannel electroglottograph*, Journal of Voice, vol. 6, no. 1, 1992, pp. 36–43.

[3] C. Hamon, E. Moulines, and F. Charpentier, *A diphone synthesis system based on time-domain prosodic modifications of speech*, in Proc. ICASSP, Glasgow, U.K., May 1989, pp. 238–241.

[4] T. Dutoit and B. Gosselin, *On the use of a hybrid harmonic/stochastic model for TTS synthesis–by–concatenation*, Speech Communication, vol. 19, no. 2, 1996, pp. 119–143.

[5] E. R. Banga, C. G. Mateo, and X. F. Salgado, *Concatenative text–to–speech synthesis based on sinusoidal modelling*, in Improvements in Speech Synthesis, J. Wiley&Sons, Chichester, 2002, pp. 52–63.

[6] Y. Stylianou, *Applying the harmonic plus noise model in concatenative speech synthesis*, IEEE Trans. Speech Audio Process., vol. 9, no. 1, 2001, pp. 21–29.

[7] W. Verhelst and M. Roelands, *An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech*, in IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 1993, pp. 554-557.

[8] J. Schoentgen, *Decomposition of vocal cycle length perturbations into vocal jitter and vocal microtremor, and comparison of their size in normophonic speakers*, Journal of Voice, vol. 17, no. 2, 2003, pp. 114–125.

[9] W. B. Kleijn, *Enhancement of coded speech by constrained optimization*, in Proc. IEEE Workshop on Speech Coding, Tsukuba, Ibaraki, Japan, 2002, pp. 163–165.

[10] J. Matoušek and J. Romportl, *Automatic pitch-synchronous phonetic segmentation*, in Proc. INTERSPEECH, Brisbane, Australia, 2008, (accepted).

[11] J. R. Bellegarda, *A novel discontinuity metric for unit selection text-to-speech synthesis*, in Proc. 5th ISCA Speech Synth. Workshop, Pittsburgh, PA, June 2004, pp. 133–138.

[12] Z. Hanzlíček and J. Matoušek, *F0 transformation within the voice conversion framework*, in Proc. INTERSPEECH, Antwerp, Belgium, 2007, pp. 1961–1964.

[13] H. Valbret, E. Moulines, and J. P. Tubach, *Voice transformation using PSOLA technique*, in Proc. ICASSP, San Francisco, CA, 1992, vol. 1, pp. 145–149.

[14] E. Nöth, A. Batliner, V. Warnke, J. Haas, M. Boros, J. Buckow, R. Huber, F. Gallwitz, M. Nutt, and H. Niemann, *On the use of prosody in automatic dialogue understanding*, Speech Commun., vol. 36, no. 1–2, 2002, pp. 45–62.

[15] T. Dutoit, *Corpus-based speech synthesis*, in Springer Handbook of Speech Processing, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., Springer Berlin, Heidelberg, 2008, ch. 21, pp. 437–455.

[16] M. Sakamoto and T. Saito, *An automatic pitch-marking method using wavelet transform*, in Proc. Int. Conf. Spoken Language Processing, Vol. 3, Beijing, China, 2000, pp. 650–653.

[17] V. N. Tuan and C. d'Alessandro, *Robust glottal closure detection using the wavelet transform*, in Proc. Eur. Conf. Speech Technology, Budapest, Hungary, Sep. 1999,pp. 2805–2808.

[18] H. Hussein and O. Jokisch, *Hybrid Electroglottograph and Speech Signal based Algorithm for Pitch Marking*, in Proc. INTERSPEECH, Antwerp, Belgium, 2007, pp. 1653–1656.

[19] M. Hagmüller and G. Kubin, *Poincaré pitch marks*, Speech Communication, vol. 48, no. 12, Dec. 2006, pp. 1650–1665.

[20] X. Huang, A. Acero, and H-W. Hon, *Spoken Language Processing: A guide to theory, algorithm, and system development*, pub. Prentice Hall PTR, 2001.

[21] M. Rothenberg, J. J. Mahshie, *Monitoring vocal fold abduction through vocal fold contract area*, Journal of Speech and Hearing Research, vol. 31, Sep. 1998, pp. 338–351.

[22] J. Matoušek, J. Romportl, D. Tihelka, and Z. Tychtl, *Recent improvements on ARTIC: Czech text-to-speech system*, in Proc. INTERSPEECH, Jeju, Korea, 2004, pp. 1933–1936.

[23] P. Boersma and D. Weenink, *Praat, software for speech analysis and synthesis*, 2005, Available from: http://www.praat.org.

[24] R. Smits and B. Yegnanarayana, *Determination of instants of significant excitation in speech using group delay function*, IEEE Trans. Speech Audio Process., vol. 3, no. 5, Sep. 1995, pp. 325–333.

[25] B. Yegnanarayana and R. Smits, *A robust method for determining instants of major excitations in voiced speech*, in Proc. ICASSP, Detroit, MI, 1995, pp. 776–779.

[26] P. S. Murthy and B. Yegnanarayana, *Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals*, IEEE Trans. Speech Audio Process., vol. 7, Nov. 1999, pp. 609–619.

[27] M. Brooks, P. A. Naylor, and J. Gudnason, *A quantitative assessment of group delay methods for identifying glottal closures in voiced speech*, IEEE Trans. Audio, Speech and Language Process., vol. 14, no. 2, March 2006, pp. 456–466.

[28] H. Strube, *Determination of the instant of glottal closure from the speech wave*, J. Acoust Soc. Amer., vol. 56, no. 5, 1974, pp. 1625–1629.

[29] J. G. McKenna, *Automatic glottal closed-phase location and analysis by Kalman filtering*, in Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland, Aug. 2001, paper 142.

[30] C. Ma, Y. Kamp, and L. F. Willems, *A Frobenius norm approach to glottal closure detection from the speech signal*, IEEE Trans. Speech Audio Process., vol. 2, no. 2, Apr. 1994, pp. 258–265.

[31] H. Hussein, M. Wolff, O. Jokisch, F. Duckhorn, G. Strecha, and R. Hoffmann, A Hybrid Speech Signal Based Algorithm for Pitch Marking Using Finite State Machines, in Proc. INTERSPEECH, Brisbane, Australia, 2008, pp. 135–138.

[32] L. Cheng-Yuan and R. J. Jyh-Shing, *A two-phase pitch marking method for TD-PSOLA synthesis*, in Proc. INTERSPEECH, Jeju, Korea, 2004, pp. 1189–1192.

[33] J–H. Chen and Y–A. Kao, *Pitch marking based on an adaptable filter and a peak–valley estimation method*, Computational Linguistics and Chinese Language Processing, vol. 6, no. 2, Feb. 2001, pp. 1–12.

[34] Wen Ding and N. Campbell, *Determining polarity of speech signals based on gradient of spurious glottal waveforms*, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 1998, pp. 857–860 vol.2.

[35] M. Legát, D. Tihelka, and J. Matoušek, *Pitch marks at peaks or valleys?*, in Lecture Notes in Artificial Inteligence, 2007, pp. 502–507.

[36] M. Legát, J. Matoušek, and D. Tihelka, *A robust multi-phase pitch-mark detection algorithm*, in Proc. INTERSPEECH, Antwerp, Belgium, 2007, pp. 1641–1644.

[37] J. Matoušek, D. Tihelka, and J. Romportl, *Building of a speech corpus optimised for unit selection TTS synthesis*, in Proc. 6th International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008.

[38] J. Matoušek, J. Psutka, and J. Krůta, *Design of speech corpus for text-to-speech synthesis*, in Proc. EUROSPEECH, Aalborg, Denmark, 2001, pp. 2047–2050.

[39] A. de Cheveigné and H. Kawahara, *YIN, a fundamental frequency estimator for speech and music*, J. Acoust. Soc. Amer., vol. 111, no. 4, 2002, pp. 1917–1930.

[40] J. M. Chambers, W. S. Kleiner, P. A. Tukey, *Graphical Methods for Data Analysis*, Wadsworth & Brooks/Cole, 1983.