A Robust Multi-Phase Pitch-Mark Detection Algorithm

Milan Legát, Jindřich Matoušek, Daniel Tihelka

Dept. of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Czech Republic

Abstract

This paper describes a robust multi-phase algorithm for marking of pitch pulses in speech using both glottal and speech signals. In the first phase, the glottal signal is used for the estimation of the fundamental frequency (f_0) contour of the given sentence. Next, pitch mark candidates are generated on the basis of both glottal and speech signals. In the third phase, the best sequence of pitch marks is found in the set of the candidates. Finally, this pitch mark sequence is post-processed. One of the features of the new method is that every pitch mark detected is given confidence, so that problematic pitch mark subsequences can be located. The algorithm was tested and compared with other pitch-mark detection methods.

Index Terms: speech synthesis, pitch pulses, glottal closure instants, pitch marks, fundamental frequency

10.21437/Interspeech.2007-457

1. Introduction

The modern pitch-synchronous methods of speech processing rely on a knowledge of the moments of glottal closure in speech signals. These moments are called glottal closure instants (GCIs), or pitch-marks (if their location in speech signal is to be found). They are generally used in pitch-synchronous speech synthesis methods (e.g. PSOLA, some kinds of sinusoidal synthesis, etc.), where they ensure that speech is synthesised in a consistent manner. Knowing the position of pitch marks, a very accurate estimation of f_0 contour could be obtained and utilised in a number of speech analysis and processing methods. Thus, it is obvious that there is a need of having a very accurate pitch-mark detection algorithm. Since the current freely available pitch-marking algorithms are error-prone (see Tab. 1), our effort was to design a robust method for a very accurate pitch-mark detection. The knowledge of pitch-mark locations is used in our speech synthesis systems [1].

The problem of pitch marking has been dealt with by several approaches, including wavelet-based analysis [2], application of nonlinear system theory [3] and many methods based on or similar to autocorrelation analysis and/or thresholding [4]. In this paper, we propose a robust algorithm based on several methods of pitch tracking and pitch pulses marking. This algorithm works in several phases and uses dynamic programming routine to find the best pitch marks sequence.

We have organised this paper into four sections. In Section 2, we briefly describe our baseline method for pitch marking and we attempt to outline the main problems of this method. Section 3 explains steps of the proposed multi-phase algorithm and the Section 4 serves to demonstrate the performance of the proposed algorithm.

This research was supported by MSMT, project No. 2C06020, and GA CR, project No. 102/06/P205.

2. The baseline algorithm

This section describes a fast and simple pitch marking algorithm [4] henceforth referred to as Baseline Algorithm (BLA) and outlines the problems of this algorithm.

2.1. Description

As glottal signal (capturing the activity of the vocal folds 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 often desirable to highlight its features with some preprocessing. In our system, the difference of the glottal signal is used at the input of the pitch marking algorithm. The difference signal emphasises the moments of glottal closures (see Fig. 1) since it reflects the rate of change of the status of the vocal folds. There are usually large sharp negative peaks in the difference signal which correspond to the moments of glottal closure. Moreover, using the difference signal, phenomena caused by improper vertical placing of the electrodes across the neck (not centred on the larynx) are also reduced. The core of the baseline algorithm is in the thresholding of the difference signal which in fact removes slow changes of the status of the vocal folds. The samples with values higher than the threshold (global r.m.s. of the signal) are the resulting pitch-mark candidates. The sequence of candidates is then post-processed and both "isolated" candidates (surrounded by regions of unvoiced speech) and candidates that are too close to each other are removed. Finally, the position of candidates is adjusted with respect to the time lag between glottal and speech signals (see Sec. 2.2). More details about the baseline pitch mark detection algorithm can be found in [4].

2.2. Tasks to tackle

As the baseline method proved to be very sensitive to the quality of the glottal signal, a more robust algorithm was to be researched. In this subsection, the main problems of the baseline algorithm are illustrated.

Fig. 1 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 this would lead to the removal of the correct peaks in the segments of low amplitude glottal signal as well. This failure causes spurious placing of pitch marks.

Another problem of the method based on difference function is demonstrated in Fig. 2. It is quite obvious that pitch marks should be placed in all energy peaks in the voiced section shown. 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.

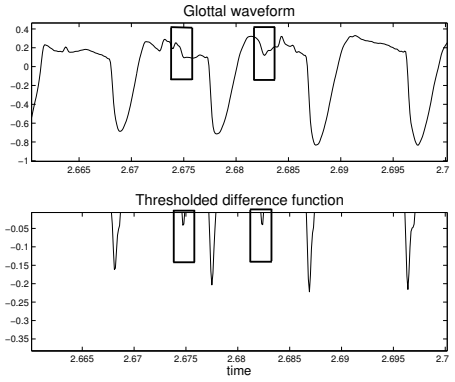


Figure 1: *False peaks in thresholded difference function.*

As stated above, there is a time lag between glottal and speech signals caused by the different places in the vocal tract where the signals are measured (larynx or lips, respectively). The length of the lag can vary with speaker (his/her vocal tract dimension), placement of the measuring equipment but it can also be affected by mastering (a process of speech signal enhancement based on filtering) of recorded signals. The typical value of the lag is about 500 μs, but mastering can cause longer time lags (see the right part of Fig. 3). Due to this fact the baseline algorithm fails.

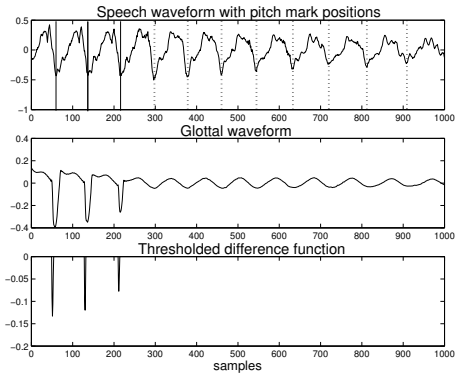


Figure 2: *Missing pitch marks are depicted by dotted lines (in speech waveform).*

3. Multi-phase algorithm description

The Multi-Phase Algorithm (MPA) proposed in this paper is based on a combination of several approaches resulting in a more robust and consistent method. The main idea is to utilise both glottal and speech signal for pitch marking.

3.1. Voicing estimation

The very first step of the proposed method is the estimation of voicing. Voicing contour is very important for marking of pulses in segments where no edges are found in glottal waveform (see Fig. 2). We use two voicing contours in further steps of the algorithm. One is obtained directly from the glottal waveform and one from the filtered speech waveform. For voicing classification we use values of short-time energy function, zero-crossing function and YIN algorithm [5] as predictors.

3.2. Fundamental frequency estimation

Having obtained the voicing contour, the next step is to estimate f_0 contour. We have decided to utilise YIN algorithm [5]

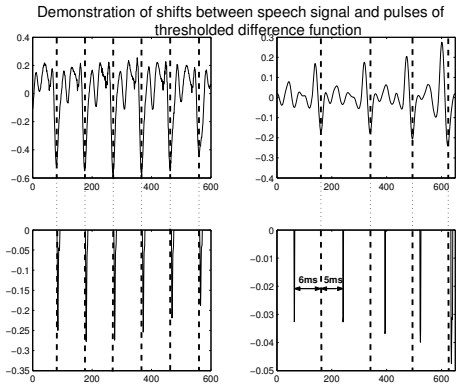


Figure 3: *Shifts between speech and glottal waveform can vary in a large range. In the left part there is almost no shift, while in the right part the shift is about 5 ms.*

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

In the frames where YIN algorithm gives uncertain results we try to use thresholded difference function. The f_0 estimate is found as an inverse of median value of differences between peaks of difference function. This estimation is used only if the number of peaks in the frame is higher than a minimum number $n_{\min}$ (the values depend on the frame length and 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.

3.3. Pitch mark candidates generation

Having estimates of both pitch and voicing contour in speech and glottal waveform, we continue with the generation of pitch mark candidates. In this phase we use two approaches to obtain a set of pitch mark candidates. For each final pitch mark three candidate positions are found in this step.

3.3.1. Modified difference function method

We use a modified BLA to generate a sequence of "difference" candidates. The modification consists in pitch-mark smoothing (see Sec. 2.1), where the estimated pitch contour is taken into account . In that way, we partially solve the problems of false peaks described in Sec. 2.2.

As there is also a problem of missing candidates (see Sec. 2.2), we use voicing contours and speech signal filtered by FIR filter with cutoff frequency 1kHz to add missing candidates into the "difference" sequence. This procedure can be summarised 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

$$[c^{(i)} + 0.8T_0^{(i)}, c^{(i)} + 1.4T_0^{(i)}], \qquad (1)$$

is explored, where $c^{(i)}$ is the last candidate in subsequence and $T_0^{(i)}$ is the local estimate of pitch period. The constants 0.8 and 1.4 were set ad hoc.

3. If there is a peak in this region whose amplitude is higher than voiced threshold thr_{v}, the position of this peak is added into the candidate sequence. This procedure is repeated until all maximum points to the right of the last candidate are found. Similarly, all predecessors of the first candidate in the subsequence can be found.

The value of constant thr_{v} depends on the voicing state of the frame in which the candidate is placed. If the frame is voiced according to speech based voicing contour, the threshold is lower than if the frame is voiced according to glottal based voicing contour, and it is much higher if the frame is marked as unvoiced in both voicing contours ($thr_{\mathrm{v}} \in \langle 0.35 * r.m.s, 1.1 * r.m.s \rangle$).

3.3.2. Simple pitch marking method

We have stated above that the final pitch mark sequence is found by dynamic programming routine in the set of candidates. At this stage we have only one candidate for each final pitch mark obtained as described in the previous subsection. We use Simple Pitch Marking Method (SPM) [6] to generate two more candidates for each final pitch mark.

This method can be summarised as follows:

1. The global maximum is found in every voiced subsequence. The position of this maximum is the first pitch mark candidate.
2. Other candidates to the right of this maximum are searched for by exploring the search region defined as

$$[c^{(i)} + \alpha T_0^{(i)}, c^{(i)} + (2-\alpha) T_0^{(i)}], \tag{2}$$

where $c^{(i)}$ is the position of maximum and $T_0^{(i)}$ is the local estimate of pitch period and α is a constant in the range of $\langle 0.5, 0.9 \rangle$.

3. In this search region two highest peaks are found. If the amplitude of the lower one is at least 80% of the amplitude of the highest one, both peaks are used as candidates. Otherwise, only the position of the highest peak is used as a "doubled" candidate (two candidates at the same position).
4. This procedure is repeated until all pitch mark candidates are found on the right side. To find all candidates to the left of the global maximum we use the same procedure, only the search region is different:

$$[c^{(i)} - (2-\alpha) T_0^{(i)}, c^{(i)} - \alpha T_0^{(i)}]. \tag{3}$$

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 two candidate positions for each final pitch mark.

3.4. Finding of the final pitch mark sequence

Before we start with searching for the final sequence, all candidates need to be given a score. In every search region (i.e. region in which one final pitch mark will be placed) there are three candidates (or one and one "doubled"). The scoring of candidates is based on two characteristics of pitch marks: the amplitude of speech signal at pitch mark position should be large and the sequence of pitch marks should correspond with the estimated f_0 contour.

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 global maximum in the search region. This measure is then normalised [6] so that the sum of amplitude scores of candidates in a given search region is equal to 1:

$$s_i(j) = \frac{\bar{s}_i(j)}{\sum_{k=1}^{3} \bar{s}_i(k)}. \tag{5}$$

The position score of two candidates is defined [6] as:

$$\bar{t}(i,j) = \frac{1}{1 + |f_0 - \frac{fs}{d}|}, \tag{6}$$

where f_0 is the local estimate of fundamental frequency; fs 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 [6], as follows :

$$t_i(j_1, j_2) = \frac{\bar{t}_i(j_1, j_2)}{\sum_{k=1}^{3} \bar{t}_i(j_1, k)}. \tag{7}$$

The final 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 a search region has three scores as it has three prospective successors (or one and one "doubled"). Once the sequence of candidates has been scored, the final sequence with the highest total score is searched for by means of dynamic programming.

3.5. Post-processing

Occasionally, there can be some errors in the final pitch mark sequence. This is because we use sequence of "difference" candidates to determine search regions for final pitch-marks.

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 (less) than $thr_{\mathrm{m}} \cdot f_0$ ($thr_{\mathrm{d}} \cdot f_0$), the missing (doubled) error is detected.

3.6. Scoring of the final sequence

The advantage of the proposed algorithm is that, along with the final pitch mark sequence, the 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 glottal waveform). If the final pitch mark is placed in the voiced speech segment according to speech based voicing contour (see 2.2), we raise its

confidence by 10%. These raising percentage values were set experimentally.

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

4. Experiments

To find out how the proposed algorithm performs in various cases and in comparison with other available algorithms, we have conducted an experiment in three languages – Czech, German (DE-M) and Slovak (SK-F). In Czech language we have tested both male (CZ-M) and female speaker (CZ-F). Besides these "standard" sentences we have tested the performance of algorithms using sentences which contain "tasks to tackle" (TT), see Sec. 2.2. Four algorithms have been tested (our baseline (BSA) and multi-phase (MPA) algorithm and also two freely available pitch marking algorithms, Edinburgh Speech Tools Library (EST, `http://www.cstr.ed.ac.uk/projects/speech_tools`) and Speech filling system (SFS, `http://www.phon.ucl.ac.uk/resource/sfs`)), see Tab. 1 for results.

SFS was tested in two ways. First, we found instants of glottal closure in glottal waveform and moved these instants into positions of energy peaks in speech waveform (SFS-G). Second, 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).

To be able to compare pitch marking methods we had to place pitch marks in test sentences manually. These sentences contain about 18.700 reference pitch marks. For comparison about 95% of these pitch marks was used, as the correct positions of these ones were quite obvious.

To compare the sequence of manually detected pitch marks (S_R) and the 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) [4]. The accuracy of automatic pitch mark detection is defined as follows:

$$Accuracy = \frac{N_R - N_S - N_D - N_I}{N_R} \times 100[\%], \quad (8)$$

where N_R is the number of pitch marks in the reference sentence S_R, N_S is the number of substitutions, N_D is the number of deletions and N_I is the number of insertions involved in the comparison process.

5. Conclusions & future work

We have presented a robust multi-phase algorithm which utilises both glottal and speech signals for pitch marking. Its performance was tested by experiments and compared with two freely available pitch marking algorithms. According to our experiments, the accuracies of the tested algorithms over various sets of sentences and laguages are as follows: EST – 88.86%; SPS-G – 87.43%; SPS-S – 55.40%; BLA – 94.69% and MPA – 97.32%. As the proposed MPA algorithm seems to be robust enough, it has been incorporated into our speech synthesis system.

Table 1: *Summary of the experiment results. "Acc" means accuracy and the values of N_D, N_I and N_S are in cases per sentence.*

CZ-M	**Acc**[%]	N_D	N_I	N_S
EST	91.44	38.6	19.8	3.6
BLA	97.66	0.4	5.8	10.2
SFS-G	97.56	4.8	5.8	7.0
SFS-S	58.27	95.6	25.2	207.4
MPA	**97.89**	4.4	3.8	7.2
CZ-F	**Acc**[%]	N_D	N_I	N_S
EST	91.20	94.8	17.8	17.6
BLA	96.08	0.6	49.8	5.8
SFS-G	92.02	13.6	63	37.2
SFS-S	67.22	131.6	29	326.2
MPA	**98.01**	8.6	7.2	12.6
DE-M	**Acc**[%]	N_D	N_I	N_S
EST	82.36	20.2	20.8	0.4
BLA	74.23	3.6	54.8	1.8
SFS-G	**94.16**	6.6	5.4	1.6
SFS-S	46.30	40.6	7.8	94.2
MPA	90.49	6	14.4	2.2
SK-F	**Acc**[%]	N_D	N_I	N_S
EST	92.78	17.6	37.4	3.4
BLA	96.07	0.6	28.4	2.6
SFS-G	70.89	38.6	58	141.8
SFS-S	40.94	69	17	427
MPA	**97.73**	5.8	8.6	3.8
TT	**Acc**[%]	N_D	N_I	N_S
EST	63.49	39.34	29	99
BLA	93.17	0.67	22	8
SFS-G	78.30	47.67	23.34	33.67
SFS-S	34.44	79.34	28.67	363.67
MPA	**97.00**	2	8	5

The future work will focus on the optimisation of the parameters involved in the MPA algorithm. In particular, we will target the search for the right parameters for German language to improve the performance of the MPA algorithm in the DE-M task. The idea is to have language oriented sets of parameters. A robust pitch-mark detection algorithm utilising speech signals only will be also researched.

6. References

[1] Matoušek, J., Romportl, J., Tihelka, D., and Tychtl, Z., "Recent Improvements on ARTIC: Czech Text-to-Speech System", Proc. Int. Conf. Spoken Language Processing, Vol. 3, Jeju, Korea, 2004, pp. 1933–1936.

[2] Sakamoto, M. and Saito, T., "An Automatic Pitch-Marking Method using Wavelet Trasform", Proc. Int. Conf. Spoken Language Processing, Vol. 3, Beijing, China, 2000, pp. 650–653.

[3] Hagmüller, M. and Kubin, G.,"Poincaré pitch marks", Speech Communication, 48(12):1650–1665, 2006.

[4] Matoušek, J., Psutka, J., and Krůta, J., "Design of Speech Corpus for Text-to-Speech Synthesis", Proc. EUROSPEECH, Aalborg, Denmark, 2001, pp. 2047–2050.

[5] de Cheveigné, A. and Kawahara, H., "YIN, a fundamental frequency estimator for speech and music", Journal Acoustical Society of America, 111(4):1917–1930, 2002.

[6] Lin, C-Y. and Jang, J-S.R., "A Two-Phase Pitch Marking Method for TD-PSOLA Synthesis", Proc. Int. Conf. Spoken Language Processing, Jeju, Korea, 2004, pp. 1189–1192.