Pitch Marks at Peaks or Valleys?[*]

Milan Legát, Daniel Tihelka, and Jindřich Matoušek

University of West Bohemia, Faculty of Applied Sciences,
Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
{legatm, dtihelka, jmatouse}@kky.zcu.cz

Abstract. This paper deals with the problem of speech waveform polarity. As the polarity of speech waveform can influence the performance of pitch marking algorithms (see Sec. 4), a simple method for the speech signal polarity determination is presented in the paper. We call this problem peak/valley decision making, i.e. making of decision whether pitch marks should be placed at peaks (local maxima) or at valleys (local minima) of a speech waveform. Besides, the proposed method can be utilized to check the polarity consistence of a speech corpus, which is important for the concatenation of speech units in speech synthesis.

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 we speak about their location in speech. They are usually used in pitch-synchronous speech synthesis methods (e.g. PSOLA or some kinds of sinusoidal synthesis), where they ensure that speech is synthesized in a consistent manner. Knowing the position of pitch marks, a very accurate estimation of f_0 contour could be obtained and utilized in a number of speech analysis and processing methods.

The problem of pitch marking has been tackled by several approaches including wavelet-based analysis [1], application of nonlinear system theory [2] and many methods based on or similar to autocorrelation analysis and/or thresholding[3]. 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 we have found out during our experiments, this decision is very important for the performance of the pitch marking algorithm in terms of its accuracy and robustness. In [4] the problem of peak/valley decision making is solved by comparing of 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.

Though this method gives quite reliable results, there are some disadvantages. First, the estimation of f_0 contour is time-consuming. Second, f_0 estimation is an error prone

[*] This work was supported by the Ministry of Education of the Czech Republic, project No. 2C06020, and the EU 6th Framework Programme IST-034434.

V. Matoušek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 502–507, 2007.

task and the errors in the contour, even if it is filtered, can affect the peak/valley decision. In this paper, we propose a simple method based on the confrontation of peaks and valleys of a speech waveform.

This paper is organized as follows. Section 2 serves to describe the proposed method. In Section 3, we briefly discuss the effects of the speech signal polarity on the synthetic speech. In Section 4, we describe our experiments to demonstrate the performance of the proposed method. Section 5 gives the conclusions of this paper.

2 The Proposed Method

2.1 Motivation

During the development of our pitch marking algorithm [5] we observed large variation in its performance. We have found out that this was due to the polarity mismatch present in our speech corpus [6]. This mismatch is illustrated in Fig. 1, where two segments of two different sentences are shown.

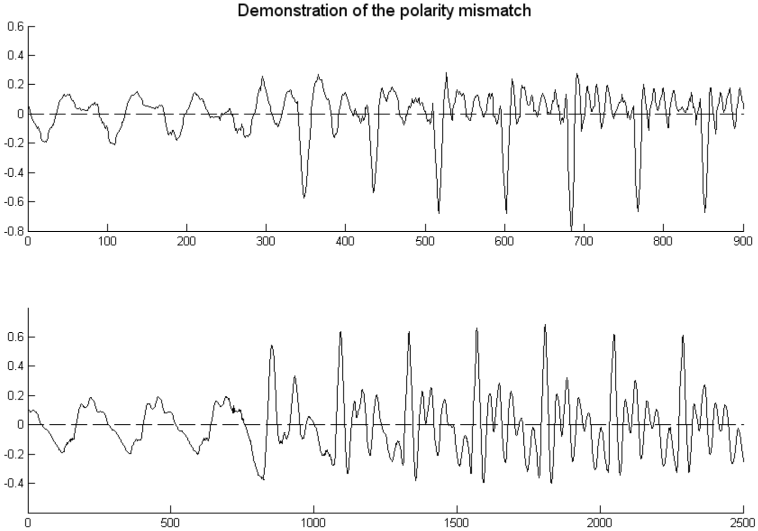

Fig. 1. Polarity mismatch. In the upper part there is a segment of the Sentence1 (negative polarity), while in the lower part there is a segment cut from the Sentence2 (positive polarity).

This observation led to the idea of development of peak/valley decision making method.

2.2 Description

Before we employ an automatic algorithm, the typical f_0 of the given speaker needs to be estimated. As the single speaker is recorded during corpus generation, this task can be very simply accomplished manually. Once we obtain the typical f_0 of the speaker, we can use this information as an input of the automatic peak/valley decision making method.

The proposed method can be summarized as follows. First, the speech waveform should be pre-processed. The aim of pre-processing is to reduce higher frequencies present in unvoiced segments and an extraneous noise. We accomplish this task by low-pass filtering by 23-rd order FIR filter with the cutoff frequency 900 Hz. The parameters of the filter were set ad hoc. This filter removes high frequencies and saves the valleys and peaks in voiced segments (See Fig. 2). The next step of pre-processing is the signal scaling. The aim of the scaling is to obtain a signal with zero mean value. This is necessary for later stages of the algorithm.

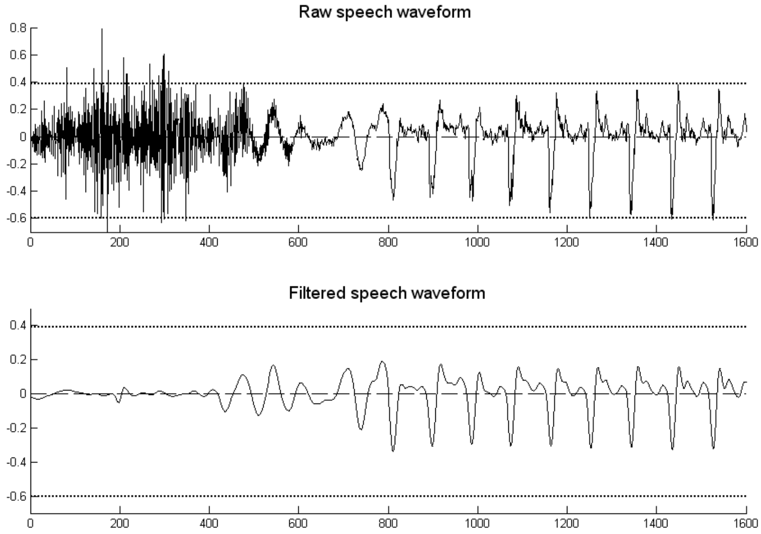

Fig. 2. Raw and filtered speech waveform. The dotted lines serve to illustrate how the noise can influence the peak/valley decision.

Having the pre-processed speech waveform, the next step of the proposed method is to confront the 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. Reset the counters *peak_count* and *valley_count*.
2. Find global maximum of the *abs_speech*. Denote its time coordinate as t_m.
3. If the position of this maximum corresponds with the position of the peak in *speech*, increment the counter *peak_count*, otherwise *valley_count* is incremented.

4. Set the value of *abs_speech* to zero in the range:
$$[t_m - 2/3 * f_0, t_m + 2/3 * f_0], \tag{2}$$
where f_0 is the estimate of speaker's typical value of the fundamental frequency. The length of this range was set experimentally.
5. Repeat steps 2, 3 and 4 until the *rms* value of the *abs_speech* is lower than $thresh * rms\_speech$, where *rms_speech* is the *rms* value of the signal *speech*. The range of the constant $thresh$ is $[0.2, 0.7]$, the higher this value is the faster the peak/valley decision is made.

In fact, the above mentioned algorithm confronts the 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*. The values of these energies are used as auxiliary predictors. If the value of the counter *peak_count* is higher than *valley_count* and *e_above* is higher than *e_below*, peaks are decided to be convenient for pitch marks placement and vice versa. If the values of counters are not in accordance with the values of energies, the decision is made solely on the basis of the values of the counters, but in this case it is marked as uncertain.

3 Discussion

In this section, we would like to address the issue of signal polarity unification during corpus recording. We have recently recorded a new speech corpus [6]. Although we placed emphasis on keeping recording conditions equal through all the recording sessions, we found several recording sessions to have speech signal with inverted polarity (See Fig. 1) – we are still examining the causes of this, but it may be related to the assembling/dissembling of the recording devices due to sharing the recording room with other projects.

The need to unify the polarity of the speech signal of all recorded phrases is obvious. As the pitch marks are placed only either at minima or maxima, phase mismatches will occur when speech units taken from signals with inverted polarity are concatenated, no matter how precisely pitch marks are detected; this is illustrated in Fig. 3. Synthetic speech will then contain audible "glitches" at such concatenation points [7], which we also confirmed in [8].

4 Experiments and Results

Rather than the experiments, in the first part of this section the results of the practical utilization of the proposed method are shown. The method was employed to check and unify the polarity of the newly recorded speech corpus [6]. The results were more than satisfactory - 98.14% of correct decisions, 1.36% of correct but uncertain decisions and only 0.5% of errors. It means that in 99.5% of cases the automatic method makes the same polarity decision as human would make. For all erroneous decisions the values of counters *peak_count* and *valley_count* were almost equal, so that these errors can be very simply detected or corrected by the setting of a threshold.

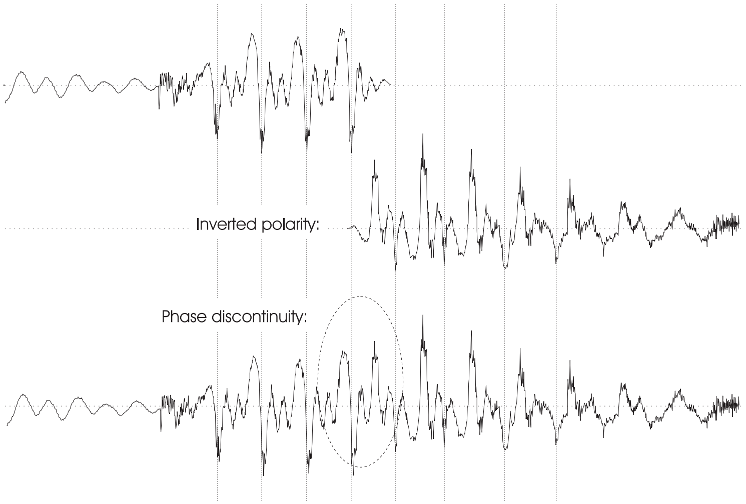

Fig. 3. Phase mismatch when units with different speech polarity are concatenated. Pitch-marks are placed at the negative amplitudes (valleys) of speech signal.

Besides, we have designed an experiment to find out how the peak/valley decision influences the performance of a pitch marking algorithm. For this purpose we have used the pitch marking method described in [5]. We have tested the performance of this algorithm depending on the type of pitch mark positions – either peaks(local maxima) or valleys (local minima) of the speech waveform. The experiment was conducted in three languages – Czech (CZ-M male and CZ-F female), Slovak (SK-F) and German (GE-M). In 8 sentences (i.e. about 7.000 pitch marks) the pitch marks were placed manually to test the performance of the automatic pitch marking method. Two reference pitch mark sets were used for testing – peak-based pitch marks and valley-based pitch marks. The summary of the results can be seen in Tab. 1. The average loss of accuracy if the pitch marks were placed into incorrect positions (i.e. placing to peaks if the polarity is negative and vice versa) was 8.6%.

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

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

5 Conclusion

In this paper, we have addressed the problem of speech waveform polarity. We have proposed a simple method for speech signal polarity checking. This method can be used for peak/valley decision making (i.e. the decision whether pitch marks should be placed in local maxima (peaks) or local minima (valleys) of the speech waveform), which is the first step before any pitch marking algorithm is employed. We have shown in our experiments how the peak/valley decision can influence the performance of the pitch marking algorithm. The decrease in the accuracy of the automatic pitch marking algorithm was 8.6% in our experiments.

Moreover, the proposed method can be used for checking of the recorded speech corpus in terms of its polarity consistence, as we have experienced the speech signal polarity mismatch during corpus recording.

References

1. Sakamoto, M., Saito, T.: An Automatic Pitch-Marking Method using Wavelet Trasform. In: Proc. INTERSPEECH, Beijing, China, vol. 3, pp. 650–653 (2000)
2. Hagmüller, M., Kubin, G.: Poincaré pitch marks. Speech Communication 48, 1650–1665 (2006)
3. Matoušek, J., Romportl, J., Tihelka, D., Tychtl, Z.: Recent Improvements on ARTIC: Czech Text-to-Speech System. In: Proc. INTERSPEECH. Jeju, Korea, pp. 1933–1936 (2004)
4. Lin, C.-Y., Roger Jang, J.-S.: A Two-Phase Pitch Marking Method for TD-PSOLA Synthesis. In: Proc. INTERSPEECH. Jeju, Korea, pp. 1189–1192 (2004)
5. Legát, M., Matoušek, J., Tihelka, D.: A Robust Multi-Phase Pitch-Mark Detection Algorithm. In: Proc. INTERSPEECH. Antwerp, Belgium (accepted, 2007)
6. Matoušek, J., Romportl, J.: On Building Phonetically and Prosodically Rich Speech Corpus for Text-to-Speech Synthesis. In: Proc. Computational Intelligence. San Francisco, U.S.A., pp. 442–447 (2006)
7. Huang, X., Acero, A., Hon, H-W.: Spoken Language Processing: a Guide to Theory. Algorithm and System Development Microsoft research. In: PTR 2001, Ch. 16, p. 829. Prentice-Hall, Englewood Cliffs (2001)
8. Tihelka, D., Matoušek, J.: The Analysis of Synthetic Speech Distortions. In: proceedings of 14th Czech–German Workshop. Prague, pp. 124–129 (2004)