A Hybrid Speech Signal Based Algorithm for Pitch Marking
Using Finite State Machines

H. Hussein, M. Wolff, O. Jokisch, F. Duckhorn, G. Strecha and R. Hoffmann

Laboratory of Acoustics and Speech Communication
Dresden University of Technology, 01062 Dresden, Germany
Hussein.Hussein@ias.et.tu-dresden.de

Abstract

Pitch marking is a major task in speech processing. Thus, an accurate detection of pitch marks (PM) is required. In this paper, we propose a hybrid method for pitch marking that combines outputs of two different speech signal based pitch marking algorithms (PMA). We use a finite state machine (FSM) to represent and combine the pitch marks. The hybrid PMA is implemented in four stages: preprocessing, alignment, selection and postprocessing. In the alignment stage, the preprocessed pitch marks are shifted to a local minimum of the speech signal and the confidence score for every pitch mark is calculated. The confidence scores are used as transition weights for the FSM. The PMA outputs are combined into a single sequence of pitch marks. The more accurate pitch marks with the highest confidence score are chosen in the selection stage. A PM reference database contains 10 minutes speech including manually adjusted PM. The evaluation results indicate that the proposed hybrid method outperforms the single PMAs but also other current state-of-the-art algorithms which have been evaluated on a second reference database containing 44 speakers.

Index Terms: speech analysis, pitch marking, FSM

1. Introduction

Pitch marking is a basic task for many applications of speech processing, for example in the speech signal analysis as well as in a text-to-speech (TTS) system basing on the Time-Domain Pitch-Synchronous Overlap-Add (TD-PSOLA) method. Therefore, a highly accurate detection of pitch marks is needed. A pitch mark (epoch) is defined as the beginning of a pitch period. Basically, it can be detected as maximum, minimum of the speech signal or Glottal Closure Instant (GCI) of the electroglottograph (EGG) signal. Glottal closure instants, i.e. the exact moments of excitation in the vocal tract during the phonation of vowels and semi-vowels, are generally chosen as the place of the marks [1]. The GCI of a EGG signal is characterized by a rapid increase after a strong minimum and it corresponds to the most negative peak of the speech signal [2].

Numerous approaches for pitch marking have been suggested over the past two decades. There are two major methods of detecting pitch marks which differ in the input signal assumed. The first category refers to algorithms which detect pitch marks directly from the speech signal [3][4][5] or from a residual speech signal [6]. The other category of algorithms is based on the EGG signal [6][7]. Hybrid methods combine such algorithms, either EGG and speech signal based pitch marking algorithms [1][8][9] or different speech signal based pitch marking algorithms, for example by AND/OR combinations of four PMA outputs [10].

The pitch marking algorithms suffer from common problems: missing pitch marks and falsely detected pitch marks [10]. These errors usually occur at vowel boundaries. For this purpose, we propose a novel hybrid method for pitch marking which combines the advantages of two speech signal based pitch marking algorithms to improve the overall performance compared to the single pitch marking algorithms. The hybrid algorithm was implemented using a finite state machine (FSM) [11] to represent and combine the pitch marks.

In Section 2, we present the detailed concept of the hybrid pitch marking algorithm. Section 3 provides an overview about database and evaluation criteria. Section 4 presents the evaluation results.

2. Hybrid Pitch Marking Algorithm

This section describes the proposed hybrid algorithm for pitch marking which combines the outputs of two speech signal based pitch marking algorithms. The algorithm is based on the alignment of pitch marks to the nearest negative peak of the speech signal and on the selection of more accurate pitch marks which have the highest confidence score. The block diagram of the hybrid pitch marking algorithm is shown in Figure 1. The hybrid algorithm contains the following components:

- Speech Signal
  - PMA1
    - Preprocessing
    - Alignment Module
    - Selection Module
    - Postprocessing
  - PM1
  - PMA2
    - Preprocessing
    - Alignment Module
    - Selection Module
    - Postprocessing
  - PM2
- Hybrid PM

Figure 1: Framework of the hybrid pitch marking algorithm
2.1. Complex Harmonic Filter Analysis (PMA1)

The first algorithm PMA1 is based on Complex Harmonic Filter Analysis (CHFA) [3]. It detects the base frequency contour (F0) of the speech signal in its first part. The CHFA calculates amplitude and phase concerning the base frequency of a speech signal. The middle frequency of a complex filter is an estimated base frequency. The phase of the complex filter output is periodic with the original fundamental frequency. The phase periodicity is used for the pitch detection and the amplitude enables the voiced/unvoiced decision. In the second part, the pitch marks are detected at the most negative peaks within a phase period. PM1 defines the output of the CHFA.

2.2. Glottal Closure Instant Detection Algorithm (PMA2)

This method of pitch marking is named Glottal-Closure-Instant Detection Algorithm (GCIDA) [4]. It is based on a multichannel approach that combines the robustness of short-time analysis with the accuracy of a wavelet-based transform. The algorithm marks the GCIs of the speech signal (most negative peaks). PM2 defines the output of the GCIDA method.

2.3. Preprocessing

Sometimes, the pitch marking algorithms detect false pitch marks in voiceless segments or during low amplitude levels. The preprocessing stage reduces these effects using an energy threshold. The speech signal is smoothed with a simple moving-average (SMA) operation to remove voiceless and very noisy segments. For a discrete-time signal $x(k)$ the smoothed signal $x_{\text{smooth}}(k)$ at time $k$, using $N$ samples ($N=20$), is computed as follows:

$$x_{\text{smooth}}(k) = \frac{1}{N} \sum_{i=0}^{N-1} x(k+i)$$

(1)

The short-time energy contour $x_{\text{ener}}(k)$ of the smoothed signal, using $N$ samples ($N=160$), is defined as (2). The threshold is calculated as 2% of the average value of energy. The pitch marks, whose energy is less than the threshold, are ignored.

$$x_{\text{ener}}(k) = \sum_{i=k-\frac{N}{2}}^{k+\frac{N}{2}} x^2(i)$$

(2)

2.4. Alignment Module

The algorithms PMA1 and PMA2 do not always detect the pitch marks at the most negative peaks. Therefore, the preprocessed pitch marks are shifted to the nearest negative peak of the speech signal (local minimum) in a search domain $M$, where $M=\pm 1$ msec. The search domain is reduced to $\pm 0.5$ msec at the boundary of voiceless segments. The confidence score will be calculated for every pitch mark. It ranges between 0 (minimum confidence) and 1 (maximum confidence). We investigated two methods to calculate the confidence score.

2.4.1. Frequency of occurrence

The confidence score of the frequency of occurrence is calculated as follows:

$$\text{Conf}_{F1}(PM_i) = \frac{\text{Number of occurrence of } PM_i}{\text{Number of combined PMAs}}$$

(3)

For a combination of two pitch marking algorithms, each pitch mark has a confidence score (probability) of 0.5. This criterion of the confidence score was used in the Recognizer Output Voting Error Reduction (ROVER) system, which combines the outputs of multiple automatic speech recognition (ASR) systems [12].

2.4.2. Distance between pitch mark and nearest negative point

This confidence score is based on the distance between pitch mark and the nearest local minimum and it is calculated as follows:

$$\text{Conf}_{F2}(PM_i) = 1 - \frac{|PM_i - \text{local minimum}|}{\text{search domain}}$$

(4)

2.5. Selection Module

The outputs of the pitch marking algorithms are combined into a single sequence of pitch marks. The single sequence of pitch marks is searched by a selection module to choose more accurate pitch marks which have the highest confidence score. We investigated two methods for this module according to the confidence score, which is used in the alignment module.

2.5.1. Selection according to the frequency of occurrence

This method selects the pitch marks which have a total confidence score of 1, i.e., it selects only pitch marks which belong to both algorithms. These common pitch marks are considered as true pitch marks.

2.5.2. Selection according to the distance between pitch mark and nearest negative point

We used a finite state machine without input and output alphabet to represent the pitch marks and to select the more accurate pitch marks. In order to reduce the processing time for the whole signal, it is divided into small segments according to pauses. The minimal length of a pause is 40 msec. The states are corresponding to a segment between two following pitch marks. Start state and end state of the FSM correspond to the segment before the first pitch mark after a pause and to the segment after the last pitch mark before a next pause. The transitions from one state to the next state are corresponding to the pitch marks from PMA1 and PMA2. The transition weights are the confidence scores of pitch marks, which were calculated according to the distance between pitch mark and nearest negative point ($\text{Conf}_{F2}(PM_i)$). Figure 2 shows the pitch marks (vertical lines) of pitch marking algorithms (CHFA and GCIDA) and the corresponding FSM graph. The best path in the FSM graph (path with maximum weight) is searched with a dynamic programming algorithm. The pitch marks are selected according to the transition weights of the best path.

2.6. Postprocessing

Some pitch marks which are aligned to the nearest negative peak are shifted to a false position especially during low amplitude levels of the signal. Therefore, we use an energy threshold to delete them (see 2.3). The threshold is calculated as $-3\%$ of the mean value of the short-time energy contour (2). If the signal amplitude is greater than this threshold, the pitch mark is removed.

If the distance between two pitch marks is less than 2 msec, one of them is deleted. The pitch mark with a greater signal amplitude is removed.
3. Experiments

3.1. Database

The first PM reference database was extracted from high-quality UK English voices which were produced within the EU project "Technology and Corpora for Speech to Speech Translation" (TC-STAR) [13]. The recordings contain read speech from several speakers with an original sampling frequency of 96 kHz (resolution 24 bit), later down-sampled to 16 kHz (resolution 16 bit).

To evaluate the performance of the pitch marking algorithms, a subset of the TC-STAR database was used as PM reference. The reference includes 10 minutes (5 minutes female + 5 minutes male speech) of speech signal based pitch marks. These pitch marks were automatically detected using CHFA in the first step. In a second step, the pitch marks were manually checked and corrected providing high accuracy.

Another PM reference was used by the European Center of Excellence on Speech Synthesis (ECESS) [14] for a gender-dependent PMA evaluation campaign. The reference contains a studio database prepared by the Ludwig Maximilians University of Munich (LMU). It contains manually marked pitch marks of 44 speakers (22 male + 22 female speakers, English and German). The speech signals of the LMU database were provided with a sampling frequency of 16 kHz and a resolution of 16 bit.

3.2. Evaluation Criteria

There are several error measures for PMAs. We use the evaluation criteria of [2][15]. These criteria were used 2006-2008 in several PMA evaluation campaigns of ECESS [14].

3.2.1. Success rate (SR%)

Correct pitch marks have the same position as the according reference pitch marks but a tolerance interval for pitch marking is allowed. The maximum tolerance interval is ± 6 samples, assuming a sampling frequency of 16 kHz and a maximum F0 of 600 Hz. The success rate (SR%) of the PMA is defined as:

\[ SR_\% = \frac{|\{ x \in \text{Test} \land (x \in \text{Ref}) \}|}{|\text{Ref}|} \times 100\% \tag{5} \]

where \(|\text{Ref}|\) represents the set of all manually corrected pitch marks (reference) and \(|\text{Test}|\) represents the set of automatically generated pitch marks. Replicated pitch marks inside the tolerance interval are not considered as correct.

3.2.2. Accuracy (ACC%)

To consider potentially inserted and deleted pitch marks, the accuracy of the PMA is calculated as follows:

\[ \text{ACC}_\% = \frac{|\text{Corr}|}{|\text{Ref}|} - |\text{Ins}| \times 100\% \tag{6} \]

The number of correct pitch marks \(|\text{Corr}|\) is defined as:

\[ |\text{Corr}| = |\text{Ref}| - |\text{Del}| \tag{7} \]

where \(|\text{Ins}|\) is the number of all inserted pitch marks and \(|\text{Del}|\) is the number of all deleted pitch marks.

4. Results

To distinguish between both hybrid pitch marking algorithms combining CHFA and GCIDA, “Hybrid (probability)” denominates the algorithm which is based on the frequency of occurrence as confidence score. “Hybrid (FSM)” specifies the other algorithm using the distances between pitch mark and local minimum as transition weights for a FSM.

Table 1 shows the pitch marking quality considering the evaluation criteria in section 3.2 for single and hybrid algorithms using the complete PM reference database from TC-STAR. With regard to success rate and accuracy, the hybrid speech signal based algorithm using an FSM outperforms the other single and hybrid algorithms (SR=96.50% and ACC=85.69% for female and male speech).

To compare the novel hybrid PMA using FSMs with other advanced algorithms, we tested “Hybrid (FSM)” in the latest
PMA evaluation campaign of ECESS in January 2008. The well-known PMA from PRAAT [16] served as reference algorithm. Additionally, three current PMAs from the ECESS members University of the Basque Country (UBC), Middle East Technical University (METU) and University of Maribor (UMB) have been evaluated. These algorithms can not be described in this paper but represent methods which were already optimized during the preceding PMA evaluation process of ECESS. Figure 3 shows the PMA evaluation results. Most of the proposed algorithms perform better on female speech. The LMU database covers a variety of speakers providing several error sources with regard to correct pitch marking. The hybrid, FSM based algorithm achieves the best overall results for male and female speakers in both categories: success rate and accuracy (SR=84.48% and ACC=67.41%).

5. Conclusion

A novel hybrid speech signal based algorithm for pitch marking using FSM was presented. The algorithm combines the outputs of two different speech signal based pitch marking algorithms.

To evaluate the algorithmic performance, we used a subset of the TC-STAR database with one female and one male speaker as pitch mark reference (10 min. speech). All reference pitch marks were automatically generated and manually corrected. Additionally, a PM reference database from Ludwig Maximilians University of Munich (LMU) was tested in a gender-dependent PMA evaluation campaign of ECESS. It contains the manually checked pitch marks from 44 speakers (22 male + 22 female). Experimental results on the PM reference database of TC-STAR indicate that the hybrid algorithm using FSM performs better than the single pitch marking algorithms. The gender-dependent PMA evaluation results on the PM reference database of LMU database show that most of the proposed algorithms perform better for female than for male speech. The novel hybrid, FSM based algorithm outperforms the other presented algorithms but nevertheless it requires further improvements in both categories: accuracy and success rate and mainly for male speech.

Considering hybrid PMA, further improvements can be obtained by combining more than two sufficient single algorithms for pitch marking which are based on preferably different concepts for filtering or detection.

6. References