ROBUST PITCH PERIOD DETECTION USING DYNAMIC PROGRAMMING WITH AN ANN COST FUNCTION

S. Harbeck *A. Kießling* *R. Kompe* *H. Niemann* *E. Nöth*
Univ. Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Inf. 5), Martensstr. 3, 91058 Erlangen, F.R. of Germany
e-mail: snharbec@informatik.uni-erlangen.de

ABSTRACT

In this paper, a new pitch synchronous F0-algorithm is described. The task of detecting pitch periods in the speech signal is solved with a search for an optimal path through a space of pitch period hypotheses. The search is efficiently implemented by dynamic programming (DP). The DP cost function is computed with an automatically trained artificial neural network (ANN) which combines the outputs of heuristic functions measuring the similarity of adjacent period hypotheses. With this algorithm a coarse error rate of 4,75% on a German speech database is achieved. It outperforms the DPF0 algorithm, which itselfs outperforms two "conventional" algorithms.

1. INTRODUCTION

Pitch is an important prosodic parameters in speech. It is used for marking e.g. focal accent, prosodic boundaries, or sentence mood (see [4], [8]). In [1] e.g. the importance of prosody and especially of pitch for spontaneous speech was proven. Because of its importance, lots of algorithms for pitch determination have already been developed (for an overview see [2]), but none of them has proven to work robustly and/or accurately for all possible circumstances. In [2] two categories of pitch determination algorithms are distinguished: *short term* algorithms compute a mean pitch frequency per frame, supposing that pitch does not change within short speech segments. In contrast to that, the category of *pitch synchronous* (also called time domain) algorithms detect in the speech signal each pitch period corresponding to the instant of glottal closure. In general, the results of the pitch synchronous algorithms are more accurate but usually less robust than the results of the short term algorithms (cf. [2]).

The results of pitch synchronous F0-algorithms can be used in different areas of speech processing, e.g. for the analysis of micro prosody, for the detection of irregular portions of speech (i.e. laryngealizations), and for pitch synchronous based feature extraction. Moreover, there is a strong interest on high–precision pitch period determination with respect to text–to–speech synthesis as e.g. for the PSOLA technique (cf. [5]).

In this paper, a new pitch synchronous algorithm is presented. It interprets the search for the pitch periods as an optimization problem. Instead of detecting single pitch periods independently from the other periods in a voiced speech segment an optimal path along pitch hypotheses with respect to an ANN cost function, which can be automatically trained, is determined.

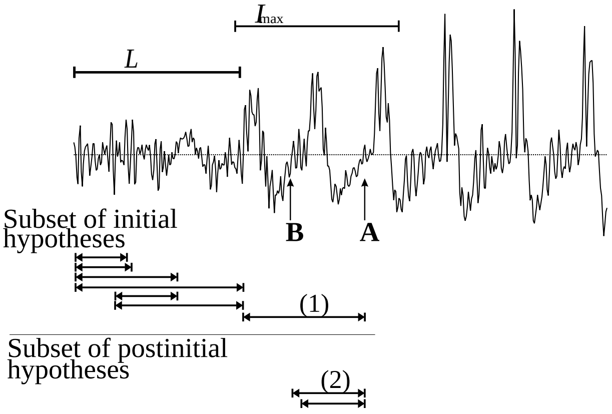
L_{max}
L
Subset of initial hypotheses
B
A
(1)
Subset of postinitial hypotheses
(2)

Figure 1. Illustration of generating initial hypotheses.

2. THE ALGORITHM

2.1. Basic Idea

Visual inspection of speech signals show a more or less periodic structure within voiced segments; they can be divided into similar looking small pieces: the pitch periods. Because excitation is only quasi–periodic pitch periods are not exactly of the same shape; as the excitation and the properties of the vocal cords and the vocal tract are changing in time, even adjacent pitch periods look more or less different. These differences are very small, especially the length of neighboring pitch periods differs usually by not more than 5 percent. For the determination of pitch periods, it makes no sense to search for them in isolation, i.e. without considering their vicinity. Thus, in the algorithm described here, the path through a search space of pitch period hypotheses is determined, which among other criteria has minimal differences between neighboring pitch periods. The search for the best path is efficiently implemented by dynamic programming (DP), an algorithm used for optimization problems, when a monotonous cost function is provided. The cost function is computed by means of an artificial neural network (ANN) that can be trained automatically and combines the outputs of heuristic functions measuring the similarity between adjacent pitch periods.

Following these ideas the PDDP algorithm (Pitch Detection with DP) has been developed. It is divided into 4 steps: after preprocessing the sampled speech signal, the positive zero crossings for each voiced segment are determined, then the search space of pitch period hypotheses is generated, and finally the best path is computed. The period hypotheses along this path represent the sequence of pitch periods within the voiced segment.

2.2. Preprocessing

In the preprocessing step the speech signal is normalized to a mean value of zero and the voiced regions are determined by an external frame-based voiced/unvoiced deci-

[1] This work was funded by the German Federal Ministry for Education, Science, Research and Technology (*BMBF*) in the framework of the Verbmobil Project under Grant 01 IV 102 H/0. The responsibility for the contents of this study lies with the authors.

Appeared in *Proc. EUROSPEECH-95* (Vol.II, pp.1337-1341)

sion (frame length: 10 ms) based on threshold relations for zero-crossing rate, signal energy, and maximum signal amplitude [3]. To restrict the search space of period hypotheses, for each voiced segment or optionally for the whole utterance an overall pitch level $\hat{G}$ is estimated with a common short term F0-algorithm like AMDF (Average magnitude difference function, cf. [6]) or DPF0 (Dynamic Programming F0, cf. [3]).

2.3. Detection and filtering of zero crossings

In PDDP a pitch period is defined as a speech segment between two positive zero crossings. Therefore, the positive zero crossings have to be extracted from the signal representing the starting points of possible period hypotheses. Each starting point is provided with additional information describing the "shape" of the signal between the starting point and the subsequent positive zero crossing (e.g. energy, length, mean, position and height of maximum and minimum). These attributes are needed for the search in the space of period hypotheses (cf. section 2.4). Idealiter, there is exactly one positive zero crossing per pitch period, but factual, a pitch period of a speech signal low pass filtered with 6,4 kHz contains many positive zero crossings. To accelerate the search, the zero crossings that are for sure not a starting point of a pitch period are therefore eliminated, following heuristic criteria that were developed by preliminary investigations on a database with reference period markers:

- If the distance to the preceding positive zero crossing is smaller than a threshold depending on the computed pitch level $\hat{G}$ and the maximum amplitude is smaller than a threshold relative to the maximum amplitude within the whole frame, then the zero crossing is not considered as a starting point and eliminated.
- If the mean of the signal values inside one segment, the region of speech between two adjacent positive zero crossings, is negative and the mean of the succeeding segment is greater, then the zero crossing is eliminated.
- Elimination of all positive zero crossings, the mean of which is smaller than the mean of the preceding positive zero crossing.

2.4. Search in the space of period hypotheses

Based on the list of positive zero crossing, the pitch level estimate $\hat{G}$ and a parameter p (the *permitted deviation* of the length of two adjacent periods), the search space of period hypotheses for a voiced region is constructed. Only pitch period hypotheses are allowed, the length of whom is restricted to the interval $[I_{min}, I_{max}]$, where

$$I_{min} = \frac{1}{\hat{G} \cdot \frac{100+p}{100}} \text{ seconds} \quad (1)$$

$$I_{max} = \frac{1}{\hat{G} \cdot \frac{100-p}{100}} \text{ seconds} \quad (2)$$

The generation of the search space of period hypotheses and the search itself is performed simultaneously. For a start, every period hypothesis, whose starting point is not later than $L = I_{max}$ from the first positive zero crossing of the voiced segment is called an *initial* hypothesis (see Figure 1; A denotes the latest endpoint of all initial hypotheses). Initial hypotheses have no predecessor and get as initial minimal costs the value 0. This strategy can cause problems especially for relatively short voiced segments: e.g. the period hypotheses (1) and (2) in Figure 1 are both ending at point A. Hypothesis (1) is an initial hypothesis with minimal cost 0 because its starting point lies within the interval L. Hypothesis (2) is a "normal" period hypothesis with costs usually greater than zero. Thus, a continuation of the path for (1) is cheaper than for (2); this means, that longer periods are prefered which may lead to subharmonic errors. To avoid this effect, the so-called *postinitial* hypotheses are introduced, which have minimal costs 0 **and** an optimal predecessor. All postinitial hypotheses have their starting point later than $L = I_{max}$ from the first positive zero crossing of the voiced segment and their endpoints are located not later than point A, which is the latest endpoint of all initial hypotheses.

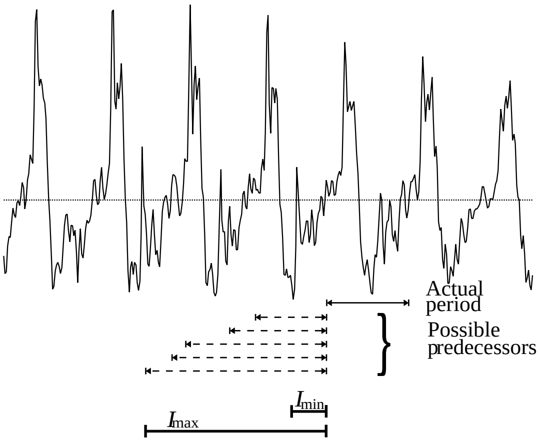

Figure 2. Possible predecessors for an actual pitch hypothesis.

For each hypotheses p there are a number of possible paths P_k leading to it with $k \in PATH(p)$ the set of indices of all paths to p. Each of them have the following form:

$$P_k = \{p_{k_0}, \ldots, p_{k_{n(k)}-1}, p_{k_{n(k)}=p}\}. \quad (3)$$

$n(k)$ is the length in number of hypotheses of the path P_k and the last hypothesis in this path is the actual hypothesis p. Each of the possible paths begins with an initial hypothesis

$$p_{k_0} \in \{\text{Initial hypotheses}\} \quad (4)$$

and

$$p_{k_j}, p_{k_{j+1}} \, \forall j = 0 \ldots n(k) - 1 \quad (5)$$

are adjacent pitch hypotheses as shown in Figure 2. The costs $C(P_k)$ of a path P_k are given by

$$C(P_k) = \sum_{l=0}^{n(k)-1} K(p_{k_l}, p_{k_{l+1}}) \quad (6)$$

with a cost function $K(.,.)$ which gets as its inputs two pitch hypotheses. The path P_j which minimizes the equation 6 is prominent under all possible paths. It contains the best path through the space of hypotheses with respect to a given cost function $K(.,.)$. Supposing a monotone cost function K the minimal costs C_p^*

$$C_p^* = \min_{k \in PATH(p)} C(P_k) \quad (7)$$

can be efficiently calculated by the following recursion

$$C_p^* = \min_{v \in PRE(p)} \left(C_v^* + K(v, p)\right), \quad (8)$$

where $PRE(p)$ is the set of possible predecessors of the hypothesis p. Supposing the minimal costs to all predecessors of p have been calculated before, the minimal costs of a path to p can be determined by choosing the

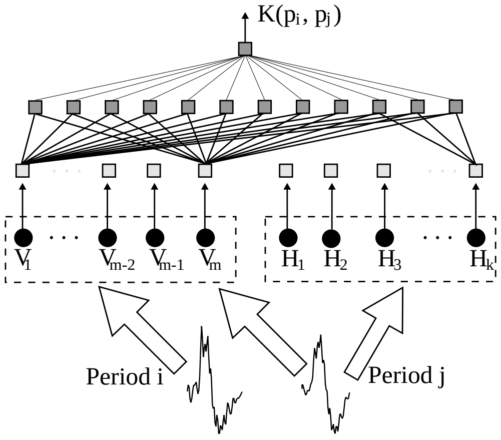

Figure 3. Calculation of the cost function $K(p_i, p_j)$ with ANN. The left group of elemental cost functions are the comparative cost functions, the right group are the elemental cost functions scoring a hypothesis itself.

best path inside all optimal paths through its predecessors. The final pitch hypotheses are determined by backtracking along the path with the minimal costs ending not more than I_{max} away from the end of the voiced segment.

2.5. Cost function

A prerequisite for a DP search is that the cost function is monotonous. Because voiced speech is usually a quasi-periodic signal with a high degree of variability, it was necessary to develop various cost functions for characterizing the similarity of two adjacent period hypotheses. In total a set of 14 so-called *primitive cost functions* was established, partially based on the attributes assigned to the zero crossings described prior. The primitive cost functions can be divided into two groups:

1. $m = 4$ (see Figure 3) cost functions characterizing a hypothesis itself (e.g. by examining the position of the global maximum or minimum inside the hypothesis)
2. $k = 10$ comparative cost functions, which describe the similarity of two adjacent period hypotheses (e.g. the nonlinear distance between two hypotheses calculated by dynamic time warping)

The results of the primitive cost functions are the input values for the ANN (see Figure 3), which combines them into one single distance value.

3. EXPERIMENTS

To evaluate the effectivity of the PDDP algorithm two error measures are defined

1. a frame based coarse error occurs if the average of all the periods within the frame differs from the reference value by more than 30 Hz
2. the period error is measured by aligning the reference period sequence and the recognized period sequence counting the shifts, deletions and insertions according to the reference period markers.

Preliminary investigations have shown that there is – as expected – a strong coherence between these two error measures. Thus, in the following only the coarse error rate is given.

The training database consists of 4 utterances of a read corpus with 1 male and 1 female speaker and in total 1 611 reference periods in 11.88 seconds of voiced speech. The evaluation database consists of 16 utterances of a spontaneous corpus recorded under different conditions from 2 male and 2 female speakers, in total 1 minute of voiced speech with 9 986 pitch periods. The databases were sampled with 16kHz and a manual pitch period marking exists for both.

Scoring function	coarse error
normalized summation	41.0 %
ANN linear activation	5.5 %
ANN sigmoid activation	5.0 %

Table 1. Coarse error rate for different scoring functions

p	# hypotheses	coarse error
20	37707	11.0 %
30	51179	8.9 %
40	78934	5.7 %
45	88886	4.8 %
50	101434	5.0 %
60	132134	6.3 %
80	273438	13.0 %

Table 2. Coarse error rate and number of hypotheses in the search space for various values of the permitted deviation (p)

Together with the information of the correct pitch periods inside the training database a training set for the ANN is constructed by calculating the values of the primitive cost functions for each correct and false pitch sequence. Because the ANN used in this algorithm has only one output node there is one target value computed for each decision. The easiest method for this target value for the ANN is to quantize it into two values: 0 for a correct sequence of pitch hypotheses and 1 for false sequences. A sequence of pitch hypotheses consists of two adjacent hypotheses, they are called "correct" when both of them are contained in the correct path of reference periods through the space of pitch hypotheses, otherwise they are called "false". A problem of this quantization is the strict and somehow unrealistic distinction between correct and false sequences, because there is a continuum including e.g. "almost correct". Therefore a second method to compute the target values was developed, which uses a heuristic quantization into six monotonous different values representing the correctness of the periods. We observed that when training an ANN with one hidden layer the way to determine the target values does not influence the results. The ANN seems to be able to learn the coherence of the primitive cost functions even with unrealistic target values.

The advantage of using an ANN for combining all 14 primitive cost functions is shown in Table 1. The summation of the (μ, σ) normalized primitive cost functions leads to a coarse error rate of 41 percent. Instead, when using a coordinate descend to get the best weights in a linear combination of the cost functions an ANN with linear activation was used. The coarse error rate of 5.5 percent was about 10 percent worse than using an ANN with sigmoid activation function. There seems to be a more complex coherence than a linear one, so in further experiments only ANNs with sigmoid activation functions are used.

The effect of filtering the positive zero crossings is shown in Table 3; the influence of various values for the permitted deviation p on the coarse error rate (with $\hat{G}$ estimated from reference, cf. below) is illustrated in Table 2. With the filtering the number of hypotheses (and therefore the

Mode	# hypotheses	coarse err.
With filtering	88886	4.8 %
Without filtering	113532	4.8 %

Table 3. Influence of the filtering of the positive zero crossing on coarse error rate and number of hypotheses in the search space ($p = 45$)

Algorithm	coarse error
PDDP with $\hat{G}$ from reference	4.8 %
PDDP with $\hat{G}$ from AMDF	6.2 %
PDDP with $\hat{G}$ from DPF0	5.3 %
DPF0	7.1 %

Table 4. Coarse error rate of the PDDP algorithm with different methods for the estimation of $\hat{G}$

Excitation	coarse error
regular	3.8 %
irregular	16.4 %

Table 5. Coarse error rate in regular and irregular speech parts

computational effort) can be reduced by 22 percent without any effect on the coarse error rate. The optimum for the permitted deviation p is approx. 45 percent, i.e. assuming that the deviation of adjacent periods is smaller than 45 percent yields the best results. When this factor is decreased the search space does not contain every correct pitch period, when it is increased even the subharmonic periods are contained in the path which are also well scored by the developed cost functions. The PDDP algorithm was also tested using different methods for estimating the pitch level $\hat{G}$ (see Table 4). Obviously the best result (4.8 %) was achieved when using the manually extracted reference pitch period markers for the estimation of $\hat{G}$ for an utterance. The worst result (6.2 %) was obtained when using the AMDF algorithm for computing $\hat{G}$. The use of the short term algorithm DPF0 [3] for estimating $\hat{G}$ leads to a error rate of 5.3 % which is about 25 % better than computing the pitch of the sentences with the frame-based DPF0 algorithm itself (7.1 %).

Because pitch detection in irregular speech (laryngealizations) is known to be a very difficult task, the PDDP-algorithm was additionally evaluated separately for regular and irregular portions of speech (see Table 5). For all the data used here a handlabelling of the laryngealized frames was available; 7.2 % of the voiced speech in the test database was marked as laryngealized. The high error rate of 16.43 % in irregular portions of speech may have two main reasons: first, there was very few training data available of pitch periods in irregular speech (too little for a robust training of the ANN) and second, the search space often did not contain the correct pitch periods due to the very long periods and the great inter-period deviations that can often be observed in laryngealizations. Other reasons for pitch detection errors even in regular parts of speech are some incorrect manual pitch period markers in the reference and the loss of pitch periods towards the end of voiced segments due to the increasing influence of unvoiced excitation.

Because the training material was very small, in further preliminary experiments PDDP was used to collect more training data with reference pitch periods in an easy and efficient way. Pitch periods in laryngograms (which were recorded in parallel to the speech signals) of 10 reread sentences of the spontaneous VERBMOBIL-corpus [9] were determined, supposing that the detection of pitch periods in laryngograms is more robust than in speech signals. The pitch periods determined in the laryngograms were mapped onto the period hypotheses in the speech signals. The periods determined in that way were then be used for training the ANN instead of the handlabelled reference periods. The same coarse error rate could be obtained, although the new training material was not manually corrected.

4. CONCLUSION

We have no possibilities to compare this PDDP with other algorithms for pitch period detection. However, since it improves even the coarse error rate of the short term algorithm DPF0 significantly (which itself outperforms other well known short term algorithms as AMDF and Seneff [7]), we can conclude to have developed a very robust algorithm. An analysis of the coarse errors showed that 30 % of the errors were located at the end or at the beginning of voiced segments, other errors occurred in laryngealizations.

The method of collecting new training material only by using the laryngogram as a basis for pitch detection and by mapping these pitch periods onto the speech signal is very encouraging. It is planned to use this method within a bootstrap strategy. To avoid the effect of an erroneous pitch reference for the construction of the training set, currently the pitch periods are extracted manually for a part of the VERBMOBIL-corpus.

In addition the use of the neural network cost function used by this algorithm as an indicator for laryngealizations is evaluated and the optimization of the search process with incorporation of a beam search algorithm are examined.

REFERENCES

[1] A. Batliner, C. Weiand, A. Kießling, and E. Nöth. Why Sentence Modality in Spontaneous Speech is more difficult to classify and why this Fact is not too bad for Prosody. In D. House and P. Touati, editors, *Proc. of an ESCA Workshop on Prosody*, pages 112–115. Lund University, Department of Linguistics, Lund, September 1993.

[2] W. Hess. *Pitch Determination of Speech Signals*, volume 3 of *Springer Series of Information Sciences*. Springer–Verlag, Berlin, 1983.

[3] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. DP-Based Determination of *F0* contours from speech signals. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages II–17–II–20, San Francisco, CA, 1992.

[4] W. Lea. Prosodic Aids to Speech Recognition. In W. Lea, editor, *Trends in Speech Recognition*, pages 166–205. Prentice–Hall Inc., Englewood Cliffs, New Jersey, 1980.

[5] E. Moulines and F. Charpentier. Pitch–Synchronous Waveform Processing Techniques for Text–to–Speech Synthesis Using Diphones. *Speech Communication*, 9(5/6):453–467, 1990.

[6] M.J. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley. Average magnitude difference function pitch extractor. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-22(5):353–362, 1974.

[7] S. Seneff. Real–time harmonic pitch detector. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-26(4):358–365, 1978.

[8] J. Vaissière. The Use of Prosodic Parameters in Automatic Speech Recognition. In H. Niemann, M. Lang, and G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, volume 46 of *NATO ASI Series F*, pages 71–99. Springer–Verlag, Berlin, 1988.

[9] W. Wahlster. Verbmobil — Translation of Face–To–Face Dialogs. In *Proc. European Conf. on Speech Communication and Technology*, volume "Opening and Plenary Sessions", pages 29–38, Berlin, September 1993.