DP-based determination of F/sub 0/contours from speech signals

Andreas Kiessling, Ralf Kompe, Heinrich Niemann, Elmar Nöth, Anton Batliner

Angaben zur Veröffentlichung / Publication details:

Kiessling, Andreas, Ralf Kompe, Heinrich Niemann, Elmar Nöth, and Anton Batliner. 1992. "DP-based determination of F/sub 0/contours from speech signals." In *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech and Signal Processing, 23 - 26 March 1992, San Francisco, CA, USA*, 17–20. New York, NY: Inst. of Electrical and Electronics Engineers. https://doi.org/10.1109/ICASSP.1992.226131.

From the Proceedings ICASSP'92.

DP-BASED DETERMINATION OF *F0* CONTOURS FROM SPEECH SIGNALS

A. Kießling, R. Kompe, H. Niemann, E. Nöth
Universität Erlangen-Nürnberg, Lehrstuhl für Informatik 5 (Mustererkennung),
Martensstr. 3, 8520 Erlangen, FRG
e-mail: kiessl@informatik.uni-erlangen.de

A. Batliner
L.M.-Universität München, Institut für Deutsche Philologie, Schellingstr. 3, 8000 München 40, FRG

ABSTRACT

A new algorithm for the determination of fundamental frequency (*F0*) contours is presented. For each voiced frame appropriate divisors of the frequency with the maximum energy in the spectrum are taken as *F0* candidates. An *F0* contour is computed using a dynamic programming (DP) method by minimizing a weighted sum of the difference between consecutive candidates and the distance of the candidates to a predetermined local target value. With this algorithm a coarse error rate of 0.6% on the frame level and of 6.4% on the sentence level is achieved on a German speech database. On the average the difference to the reference is 1.9 *Hz*. Our algorithm outperforms two "conventional" algorithms tested on the same data.

1 INTRODUCTION

The speech group at the institute in Erlangen is working on a dialog system for understanding and answering spoken queries to a train schedule database [1]. In the framework of our system we want to use prosodic information mainly for the determination of sentence mood, focal accent (and thereby focus), and boundaries within an utterance[1]. It has been shown that prosodic information is an important cue to indicate these properties of an utterance [2, 3, 4]. The most important parameter carrying prosodical information is pitch [5, 2, 6, 7]. The acoustical correlate of pitch is the fundamental frequency (*F0*)[2]. Many algorithms already have been proposed for pitch determination (for an overview see [9]). A main drawback of them is that *F0* is computed locally without taking into account information about the *F0* in other portions of the utterance. Therefore these methods usually work well in regular portions of speech, but often fail in irregular portions.

In this paper we propose a new algorithm for the determination of pitch contours of speech signals. The basic idea of this algorithm is the following: The frequency with the maximum amplitude in the short time spectrum is *F0* or a higher harmonic (*i.e.*, an integer multiple) of *F0*. For each frame appropriate divisors of this frequency are taken as *F0* candidates. To restrict the search only candidates within a certain range around a global *F0* value (pitch level) are permitted. An *F0* contour is computed using a dynamic programming (DP) method by minimizing a weighted sum of the difference between consecutive candidates and the distance of the candidates to a local target value. The algorithm is robust even when it encounters irregular portions of speech, and performs well with telephone quality speech.

2 THE ALGORITHM

Prior to the actual computation of the *F0* contour, a voiced/unvoiced classification has to be performed. We used a method after [10]: For each frame (in our case 12.5 *msec*) certain features (zero-crossing rate, signal energy, maximum signal amplitude) are computed. Based on threshold relations the frame is classified as voiced or unvoiced. Consecutive frames of the same voicing type constitute one region. Very small regions are eliminated.

For each voiced region a pitch contour (one *F0* value per frame) is computed from the low pass filtered signal (in our case at 1100 *Hz*) with the following algorithm [11]: The core of the algorithm is a DP search, which is based on the observation that *F0* changes between consecutive frames are usually small. To guide the DP search one (local) *F0* target value (G) is computed for each voiced region. This is done using a multi channel method combining different traditional algorithms. This value is computed at the frame with the maximum energy under the assumption that the *F0* determination is most reliable at this frame. In the current implementation the combined algorithms are the *AMDF*

[1] Such boundaries can be caused by prosodical phrasing or hesitations.

[2] For the relevance of other prosodic parameters, *e.g.*, intensity and duration, see [8, 6].

algorithm (a time domain correlation method [12]), and the *Seneff* algorithm (a frequency domain method [13]). The local *F0* target value is the median of values computed with these algorithms using different parameters and analysis window sizes.

The average A of all the local target values of one utterance is considered as an indicator of the pitch level of the speaker[3]. It is used to determine the *F0* candidates for each voiced frame t. First the frequency F_t with the maximum energy in the spectrum of each frame is determined. Then an integer $k_t = F_t/A$ is computed for each frame. *F0* candidates $K_{t,j}$ are computed according to the following formula:

$$K_{t,j} = \begin{cases} \frac{F_t}{k_t+j} & \text{if } k_t + j > 0 \\ \mathit{undefined} & \text{otherwise} \end{cases} \quad , \quad j \,\epsilon\, [-n, +n]$$

In our experiments a maximum of 5 candidates per frame was computed (*i.e.*, $n = 2$). This was found to be sufficient in order to have the *F0* value always in the set of candidates. Now for each voiced region an *F0* contour is searched in the matrix $K_{t,j}$ using DP. The contour is optimal with respect to the cost function $\mathcal{C}$, where S denotes the startframe, E the endframe of the region, $V(t,j)$ the preceding point in the Matrix (with respect to the cost function), and γ a weight factor whose value depends on the length of the voiced region:

$$\begin{array}{lcl}
\mathcal{C}_{S,j} & = & 0 \\
\mathcal{C}_{S+1,j} & = & |\log \frac{G}{K_{S,j}}| \\
\mathcal{C}_{t,j} & = & \gamma \cdot |\log \frac{G}{K_{t,j}}| + |\log \frac{K_{t-1,V(t,j)}}{K_{t,j}}| \\
 & & + \, \mathcal{C}_{t-1,V(t,j)} \ , \qquad S + 1 < t < E \\
\mathcal{C}_{E,j} & = & \mathcal{C}_{E-1,V(E,j)} \\
 & & \\
V(t,j) & = & \alpha, \qquad \text{where } \alpha \text{ holds:} \\
 & & \quad \min_{-n \le \alpha \le n} (\mathcal{C}_{t-1,\alpha} + |\log \frac{K_{t-1,\alpha}}{K_{t,j}}|) \\
V(E,j) & = & \alpha, \qquad \text{where } \alpha \text{ holds:} \\
 & & \quad \min_{-n \le \alpha \le n} (|\log \frac{K_{E-1,\alpha}}{K_{E,j}}|)
\end{array}$$

After some trials we set $\gamma = 0.01$ for voiced regions of more than 5 frames and $\gamma = 0.1$ else. In the case of long voiced regions it is important that the influence of the local target value during the contour search is kept very small to allow the contour to follow a great *F0* fall or rise. The logarithm of the frequencies is probably perceptually more adequate than the frequencies themselves[4]. The final *F0* contour is determined by backtracking along the path with the minimal costs. Due to the frequent occurrence of irregularities at the boundary frames of each region the search is actually performed between the frames $S+1$ and $E-1$ and the optimal path is appropriately prolongued to the frames S, and E.

3 EXPERIMENTS

For the evaluation of the algorithm for voiced/unvoiced decision a German speech database of 264 utterances (about 10 minutes of speech, from 12 speakers) was used. The sampling rate was 10 *kHz*. A manual phone labeling existed for this database. The evaluation showed that 98.8% of the frames of the pure voiced phones (*i.e.*, vowels, glides, nasals, liquids), 41.0% of the mixed frames (*i.e.*, voiced fricatives and voiced plosives), and 8.8% of the pure unvoiced frames were classified as voiced. The algorithm was tuned in order to produce more unvoiced to voiced classification errors than vice versa.

The database used for the evaluation of the voiced/unvoiced decision was not very interesting for the evaluation of the $F0$ algorithm, because the utterances were spoken with a relatively monotonous voice, therfore our algorithm for *F0* extraction was tested on two different German speech databases (databases A and B)[5]. The corpora were designed for a research project on German intonation [7]. They contained minimal sentence pairs, *i.e.*, sentence mood and focus of the second sentence was determined by the first (context) sentence. Sentence mood and focus of the second sentence could only be discriminated by intonation, *e.g.*, "Who drinks? <u>Leo</u> drinks.", "What does Leo do? Leo <u>drinks</u>.","What did you say? Leo <u>drinks</u>?" (The focus is underlined). This design of the sentences resulted in high variations of *F0* thus making them interesting for testing *F0* algorithms: For both databases the average difference between the minimal and the maximal *F0* within an utterance was about 120 *Hz*. The minimal observed *F0* was at 58 *Hz* whereas the maximum *F0* in the data was at 520 *Hz*. (These values were computed on the automatically determined and hand-corrected $F0$ contours.) Hence *F0* candidates have been limited to the interval of 55 - 550 *Hz*. It is worth mentioning that this is quite a large range. Database A consisted of 195 utterances from 7 speakers (4 male, 3 female). Database B consisted of 357 utterances from the speakers of database A except one male speaker. The signals of both databases were sampled with 16 *kHz*. With the algorithm for voiced/unvoiced decision mentioned above in database A 333 *sec* of speech were classified as voiced, in database B 469 *sec* were classified as voiced. For the *F0* determination the signals were further low pass filtered at 1100 *Hz* and downsampled by a factor of 7. *F0* values were computed for every voiced frame. Parameters and thresholds of the algorithm have been manually adjusted using database A. Database B was only used for a final test.

Before evaluating the errors of the algorithm two error measures had to be defined. Within our speech system *F0* contours will be used for determining the sentence mood and focus of utterances as well as for phrase boundary detection. For these tasks it is im-

[3] For a real-time system instead of the complete utterance a certain portion in the past has to be used.

[4] It is not yet clear if $F0$ measured in Hz or in semi-tones is more adequate [14]. In the experiments with this algorithm the logarithmic values yielded better results than the ones in Hz.

[5] Both databases were recorded at the *Institut für Phonetik* at the Ludwig-Maximilian Universität, München.

Database	*DP*	DP_s	$AMDF_s$	$Seneff_s$
A	1.7	1.6	1.9	1.7
B	0.6	0.6	1.9	1.3

Table 1: *Percentage of frames with coarse error (difference > 30 Hz to the reference).*

Database	*DP*	DP_s	$AMDF_s$	$Seneff_s$
A	12.3	8.9	30.3	17.9
B	8.1	6.4	41.9	27.9

Table 2: *Percentage of sentences with at least one coarse error.*

Database	*DP*	*AMDF*	*Seneff*
C	1.9	2.2	3.3

Table 3: *Fine errors given as mean of the absolute value of the relative difference between the reference value and the automatically computed value in Hz. (Computed only at frames without coarse error.)*

portant to have a reliable *F0* contour where the values do not have to be very accurate. Hence we define two error measures[6]: if the automatically determined *F0* value and the reference value differ by more than 30 *Hz* a **coarse error** occurred. This threshold was set heuristically after having examined database *A*. **Fine errors** are measured on all frames which don't have a coarse error. These errors are defined as the difference between the reference value and the automatically computed value in *Hz* divided by the reference value[7]. For our application the percentage of frames with coarse errors is a more interesting measure than the arithmetic value of the error itself.

Tables 1 and 2 show the coarse error rates of our algorithm (column *DP*), and for comparison error rates of the above mentioned algorithms *AMDF* and *Seneff*. The contours were smoothed using first a 3 point median and second a 5 point median (columns $DP_s, AMDF_s, Seneff_s$). Coarse errors for the unsmoothed contours of *AMDF* or *Seneff* were considerably worse and are not given in the table. Table 1 gives the percentage of frames which have a coarse error. In our application an utterance might not be analyzed correctly if at least one coarse error occurs. Thus it is more interesting to determine how many utterances have errorfree *F0* contours. Table 2 shows the percentage of utterances containing at least one coarse error. The error rates were determined by comparing the automatically computed *F0* contours manually with contours produced by a mechanical pitch detector. If necessary an exact reference value was determined from the signal and with perception tests. The fact that the performance of our algorithm on database *B* is better than on the "training" database *A* is due to the greater number of laryngealizations in database *A* (see below).

Table 3 gives the mean of the absolute values of the fine errors for the unsmoothed contours computed with the three algorithms. These errors were determined for a subset *C* of 24 sentences from database *B* and only at frames without coarse error. For this subset period-synchronized *F0* reference contours were produced semi-automatically. The fine error of 1.9 *Hz* for our algorithm is slightly better than the one of the two other algorithms.

4 DISCUSSION

We described an algorithm for detecting *F0* contours. Only about 1% of the frames and about 7% of the sentences have coarse errors.

An analysis of the coarse errors of the unsmoothed *DP* contours showed that one third of the errors were caused by a wrong computation of the target value *G*. These shall not be discussed any further, because the (conventional) algorithms for their computation are exchangeable. The other errors are due to a wrong computation of the contour itself. Most of them occurred at laryngealizations. These are irregular portions of a voiced signal or regular portions which are characterized by an extremely low *F0* [16, 10]. Frequently the transition to and from this low *F0* is abrupt. Usually our algorithm smoothes over irregular portions of a signal due to the implicit use of context information. This is perceptually adequate (see below). In rare cases it is not able to smooth: Since at laryngealizations the frequency maximum used for the computation of the *F0* candidates is sometimes extremly low, it causes the cheapest path to be too low even in the regular portions of the same voiced region. At the moment this is the main draw-back of our algorithm. Due to a relatively small irregularity the contour might be wrong at a larger portion of the utterance. This explains why our algorithm compared to *Seneff*, and *AMDF* has a low percentage of sentences with coarse error (see Table 2) opposed to only a slightly better coarse error rate on the frame level (see Table 1).

5 FUTURE WORK

Laryngealizations often cause problems in the *F0* computation, but even if the low *F0* (which can be seen easily in the visualized signal) is computed correctly this

[6] Although the distinction between the two error measures is not uniquely defined, it is well known from the literature (see for example [9, 15]).

[7] An analysis window contains more than one pitch period. Therefore a reference value is the average of all the pitch periods within the window.

does not correspond at all to the perceived pitch contour. Human listeners perceive a pitch contour which can be described as a extrapolation of the pitch of the context. This perceived pitch contour seems to be superimposed by a jar. The perceived pitch is relevant for the determination of focal accent and sentence mood. Thus a new version of the algorithm is currently implemented which first localizes laryngealizations in the voiced regions, excludes them from the *F0* computation, and extrapolates the $F0$ contour at these parts from the context.

On the other hand laryngealizations are not only causing problems but carry important information [17]: They often cooccur with the pitch fall at the end of affirmative propositions, at the beginning or end of phonation as well as at morpho-syntactic boundaries like phrase, word or morphem boundaries (*e.g.*, *see easily* vs. *cease*). The superimposed jar allows human listeners to use this information. Hence, the knowledge about the locations of the detected laryngealizations should not be discarded but should be made available to other modules of the speech understanding process.

6 ACKNOWLEDGEMENTS

This work was supported by the German Ministry for Research and Technology (*BMFT*) and by the *Deutsche Forschungsgemeinschaft* (*DFG*). Only the authors are responsible for the contents of this paper.

References

[1] H. Niemann, G. Sagerer, U. Ehrlich, E.G. Schukat-Talamazzini, and F. Kummert. The interaction of word recognition and linguistic processing in speech understanding. In P. Laface and R. de Mori, editors, *Recent Advances in Speech and Language Modeling*, Springer Verlag, Berlin, Heidelberg, New York, 1991.

[2] 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*, pages 71–99, Springer Verlag, Berlin, Heidelberg, New York, 1988.

[3] A. Batliner and E. Nöth. The prediction of focus. In *European Conf. on Speech Communication and Technology*, pages 210–213, 1989.

[4] E. Nöth. *Prosodische Information in der automatischen Spracherkennung — Berechnung und Anwendung.* Niemeyer, Tübingen, 1991.

[5] I. Lehiste. *Suprasegmentels.* MIT Press, Cambridge, Massachusetts, 1970.

[6] E. Nöth, A. Batliner, T. Kuhn, and G. Stallwitz. Intensity as a predictor of focal accent. In *XIIth Int. Cong. of Phonetic Sciences*, 1991.

[7] H. Altmann, A. Batliner, and W. Oppenrieder, editors. *Zur Intonation von Modus und Fokus im Deutschen.* Max Niemeyer Verlag, Tübingen, 1989.

[8] M. Beckmann. *Stress and Non-stress Accent.* Foris Publications, Dordrecht, 1986.

[9] W. Hess. *Pitch Determination of Speech Signals.* Volume 3 of *Springer Series of Information Sciences*, Springer Verlag, Berlin, Heidelberg, New York, 1983.

[10] A. Kießling. *Optimierung des DPGF-Grundfrequenzverfahrens durch besondere Berücksichtigung irregulärer Signalbereiche, Diplomarbeit.* Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg, 1990.

[11] R. Kompe. *Ein Mehrkanalverfahren zur Berechnung der Grundfrequenzkontur unter Einsatz der Dynamischen Programmierung, Diplomarbeit.* Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg, 1989.

[12] M. 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.

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

[14] D.J. Hermes and J.C. van Gestel. The frequency scale of speech intonation. *J. of the Acoustic Society of America*, 97–102, 1990.

[15] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. McGonegal. A comparative study of several pitch detection algorithms. *IEEE Trans. on Acoustics, Speech and Signal Processing*, ASSP-24:399–413, 1976.

[16] P. Hedelein and D. Huber. Pitch period determination of aperiodic speech signals. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 361–364, 1990.

[17] A. Kießling, R. Kompe, E. Nöth, and A. Batliner. Irregularitäten im Sprachsignal – störend oder informativ? In R. Hoffmann, editor, *Elektronische Signalverarbeitung, Studientexte zur Sprachkommunikation, Heft 8*, pages 104–108, Technische Universität Dresden, Dresden, 1991.