PITCH PERIOD ESTIMATION BY FILTERING THE FUNDAMENTAL FREQUENCY OUT OF THE SPEECH WAVEFORM

Dick R. van Bergem

1. INTRODUCTION

In the past decades a lot of algorithms for estimating the fundamental frequency of speech signals have been proposed. For an overview of the most important ones see Hess (1983). The algorithms either work in the time domain or in the spectral domain. Those working in the spectral domain use a short-term analysis of the speech signal. That is, the average pitch period duration is estimated for (usually overlapping) frames that contain a small number of successive pitch periods. Those working in the time domain estimate a train of laryngeal pulses, which are called pitch markers.
Time domain methods have some clear advantages over spectral methods if they work properly. In the first place, they are exact in locating pitch periods, whereas spectral methods can only give estimates of pitch contours and may have difficulty with frames containing both voiced and unvoiced speech. In the second place, a pitch contour derived from a series of pitch markers can be very easily checked by a careful examination of the markers. It can also easily be hand-corrected by removing or adding a number of markers. In the third place, time domain methods can (contrary to spectral methods) be used for pitch synchronous techniques such as pitch synchronous Fourier transformations or the PSOLA-technique with which the prosodic features of speech can be manipulated (Hamon et al., 1989; Charpentier and Moulines, 1989).
Recently an interesting time domain pitch extraction method was proposed by Dologlou and Carayannis (1989). They filtered the fundamental frequency out of the (digitized) speech waveform with the iterative use of a lowpass filter (a 3-point Hanning window). Comparing the filtered fundamental frequency with the output signal of a laryngograph, they observed that the points of closure of the vocal chords were very well estimated by the drops in the fundamental frequency signal that were extracted by their method.
The crucial part of their algorithm concerns the criterion with which the iterative filtering is terminated. This is done by checking at the end of each iteration whether one or more frequency components are still present in the filtered signal. For this purpose Dologlou and Carayannis use the autocorrelation function and a 2-order LPC analysis which give the same estimate of the frequency of a pure sinusoid; when the difference between these two estimates falls below a predetermined threshold the algorithm stops. However, the assumption that the fundamental frequency is a pure sinusoid is of course not correct, because it is continuously changing. Even within a small stretch of speech (Dologlou and Carayannis used non-overlapping frames of 100 ms) the fundamental frequency varies, so that the threshold they mention should be rather conservative. On the other hand, this might lead to a preliminary ending of the iterations. The problem of choosing a proper threshold is illustrated by the fact that the authors fail to mention which threshold they propose.

Another drawback of their method is the inefficient way of filtering the speech signal. A large number of iterations are required to attenuate the higher harmonics of the fundamental frequency, especially for speakers with a low pitch. This can be easily seen in the following way. The frequency response of a Hanning window is given by the formula:

$$H(\omega) = \cos^2(\omega T/2) \qquad (1)$$

in which T denotes the sampling interval. After n iterations the frequency response will be:

$$H(\omega) = \cos^{2n}(\omega T/2) \qquad (2)$$

The relative attenuation R occurring between two frequency components equals:

$$R = \left(\frac{\cos(\omega_1 T/2)}{\cos(\omega_2 T/2)}\right)^{2n} \qquad (3)$$

With formula (3) it can be verified that for a fundamental frequency of for instance 100 Hz, it takes 1554 iterations (at a sampling frequency of 10000 Hz) to attenuate the component at 200 Hz by 40 dB. For a 3-point Hanning window that looks 1 sample forward and 1 sample backward the actual window size becomes 3109 coefficients after 1554 iterations. Compare this with a lowpass Kaiser window with similar (in fact even better) specifications that requires only 225 coefficients (Crochiere and Rabiner, 1983). Apart from the large computational effort involved in the filtering process each iteration also requires the calculation of the autocorrelation function and a 2-order LPC analysis to test whether the stop criterion has been reached.
An additional problem is that the iterative filtering doesn't provide a clear passband, because it gives a monotonically decreasing frequency response. This means that the fundamental frequency itself in the above mentioned example (100 Hz) is already attenuated by almost 7 dB after 1554 iterations, which can be verified with formula (2).
In this paper an alternative to the method of Dologlou and Carayannis is presented that consists of two successive steps. First a global estimate of the pitch contour in the speech signal is made and subsequently the speech is bandpass filtered around this measured contour with a narrow passband. The use of a bandpass filter instead of a lowpass filter has the advantage of eliminating low frequency hum and noise especially for speakers with a high pitch (The experimental data of Dologlou and Carayannis show a lot of low frequency noise, although the source of this noise is rather mysterious in my view).
In chapter 2 and 3 the two stages of the algorithm will be discussed in detail. The performance of the algorithm is the subject of chapter 4 and chapter 5 gives the conclusions.

2. ESTIMATION OF THE GLOBAL PITCH CONTOUR

In principle any method can be used for our purpose to estimate the global pitch contour of a (digitized) speech waveform. An autocorrelation analysis was chosen, which is rather robust and reliable (Rabiner et al., 1976). A fixed analysis frame of 30 ms is

used and a step size of 10 ms. To avoid the unnecessary computation of the autocorrelation function a simple voiced-unvoiced classifier is used prior to the analysis of each frame (Atal and Rabiner, 1976). This classifier uses three features to label it as either voiced or unvoiced:

1. The energy of the signal in dB, after the signal level has been scaled on a long-term basis to a maximum level of e.g. 2048 (12 bits).

2. The normalized autocorrelation coefficient at unit sample delay.

3. The number of zero-crossings in the frame (the offset in the signal should be removed).

The mean values of these parameters for voiced as well as unvoiced speech and the corresponding variance-covariance matrices were estimated with the aid of 6 training sentences (2 female speakers, 4 male speakers) in which voiced and unvoiced parts had been labelled by hand. By measuring the same three parameters in each frame, we can calculate the Mahalanobis distance between the measured values and the reference data and label the frame as either voiced or unvoiced on the basis of the smallest distance.
If a frame is labelled as voiced, a compressed center clipper is used to spectrally flatten the speech (Rabiner, 1977). The clipping level is set to 80 % of the smaller of the maximum absolute signal level over the first and last one-thirds of the analysis frame. An additional advantage of the center clipper is that a large number of multiplications necessary to calculate the autocorrelations can be skipped (namely for those sample values that fall below the clipping level).
Subsequently all autocorrelations corresponding to a fundamental frequency between 60 Hz and 1000 Hz are calculated and the position and amplitude of the (normalized) peak value in the autocorrelation function is determined. After all frames have been analyzed a global pitch contour is obtained in the following steps:

1. Calculate the median of the fundamental frequency of all voiced frames.

2. Change the label of frames with a fundamental frequency higher than twice the median or lower than half the median from voiced to unvoiced.

3. Estimate the fundamental frequency of successive 100 ms segments of speech by taking the median of the 10 segment frames together with 2 preceding and 2 following frames. The peak value in the autocorrelation function can be used as an indication of the reliability of the fundamental frequency measurements.

4. Supply each unvoiced segment with the fundamental frequency value of the preceding frame (if that frame is voiced), or the fundamental frequency value of the following frame (if that frame is voiced). If both the preceding and the following frame are also unvoiced, the unvoiced segment is filled with the overall median value of the fundamental frequency.

This procedure results in a very reliable estimate of the global pitch contour, even if a great number of the individual frames are wrong estimates of the local pitch. Note that also unvoiced speech parts are given a fundamental frequency, because they may contain a few pitch periods that have been 'overlooked' by the autocorrelation analysis.

3. ESTIMATION OF PITCH MARKERS

After the global pitch contour has been established, the speech waveform is bandpass filtered around this contour. The actual filter is a Kaiser window (Crochiere and Rabiner, 1983) with the following specifications:

1. The center frequency is the fundamental frequency of the current 100 ms segment.

2. The passband is half of the center frequency.

3. The transition band is half of the center frequency.

4. Attenuation in the stopband is 40 dB.

Within a 100 ms segment the fundamental frequency will usually vary. The passband of the filter should therefore be as large as possible, so that all variations in fundamental frequency for the 100 ms are passed. On the other hand, the passband should not be too large, because higher harmonics of the varying fundamental frequency should fall outside the passband. If the passband of the filter is $pF0_{med}$ (that is, a proportion of the median of the fundamental frequency for the 100 ms segment), then the lower bound of the passband is $F0_{med} - pF0_{med}/2$. The second harmonic of this frequency should be higher than the upper bound of the passband:

$$2\,(F0_{med} - pF0_{med}/2) > F0_{med} + pF0_{med}/2 \qquad (4)$$

Elaboration of (4) gives:

$$p < 2/3 \qquad (5)$$

If we choose $p = 1/2$, the second harmonic of the lower frequency bound of the passband falls well within the transition band.
The width of the transition band is not critical. In order to get steep slopes in the filter characteristics, the transition band should be small. On the other hand, the number of filter coefficients increases with a smaller transition band. A transition band of half the median fundamental frequency appeared to be a reasonable compromise.
The number of filter coefficients is established using the overall median value for the fundamental frequency and is applied to all 100 ms segments of speech. For each segment filter coefficients are recalculated according to the current fundamental frequency segment value. To avoid discontinuities in the filtered signal the filtering of speech samples at the end of a segment is continued until the filtered signal reaches the zero line.
After the filtering has been completed pitch markers are placed at drops in the filtered signal if a certain threshold is exceeded. The drops in the filtered signal are defined as

points that have a lower sample value than the two nearest neighbours on the left side and the right side. The threshold to indicate whether a drop belongs to a voiced speech part or an unvoiced speech part must be empirically established. It was found that good results can be obtained for a threshold value at about 5% of the maximum value in the filtered signal.
If the pitch contour is desired, it can be calculated from the distances between the markers. Unvoiced parts are indicated by very large distances between markers (for instance distances that are greater than the pitch period of half the overall median fundamental frequency value may be used as a criterium). In order to get pitch values at equidistant points in time similar to the frames in spectral domain methods, a fixed stepsize can be chosen at about the pitch period length of the overall median fundamental frequency value. In this way the steps will most of the time go from one pitch period to the next, so that an optimum resolution in the pitch contour is obtained. A non-linear smoother (Rabiner & Schafer, 1978) may be used to get a smooth pitch contour.

4. RESULTS

The new algorithm was implemented in Fortran on a μVAXII minicomputer. At our Institute a few sentences were available that contained hand-labelled pitch markers (placed at zero-crossings at the beginning of prominent period peaks) which were used to change the prosodic features of the sentences with the aid of the PSOLA-technique (Hamon et al., 1989; Charpentier and Moulines, 1989). These sentences had been recorded in an anechoic room with a high quality microphone (Sennheiser MKH 105T) and a Panasonic NV-F70HQ videorecorder. The same prosodic manipulations were performed using pitch markers generated by the new algorithm and results were compared with the hand-labelled versions. It appeared that the versions based on the new algorithm gave a brighter sound than the hand-labelled versions. Perhaps this is due to the natural position at which the pitch markers are placed, namely at the point where the vocal chords close (Dologlou and Carayannis, 1989). However, a better explanation may be that the automatic procedure places the markers with more consistency than a human experimenter.
After this first test the algorithm was again used to place pitch markers in 40 sentences uttered by a male subject that were used for an experiment with PSOLA-manipulated speech (Laan, 1990). These sentences had been recorded in the same manner as the earlier mentioned ones. All markers were placed accurately apart from a few occasional errors caused by a not completely attenuated second harmonic or at a voiced-unvoiced boundary where a drop in the output signal didn't reach the threshold criterium contrary to neighbouring drops. Once the algorithm failed at the end of a vowel, when the pitch periods abruptly became about twice as long as they were (creaky voice). These occasional errors were easily traced by looking at the pitch markers in the speech signal and could also be easily corrected.
In figure 1 two examples are given of the output of the algorithm for the words /mɑn/ and /sɪs/ uttered by a man to illustrate the positions at which the pitch markers are placed. Note that the filtered output is close to zero for the unvoiced parts of the word /sɪs/. Next the performance of the new algorithm was compared with that of two other pitch detectors that are available at our Institute. The first one designed by Reetz (1989) calculates the sum of all sample values between each pair of successive zero-crossings

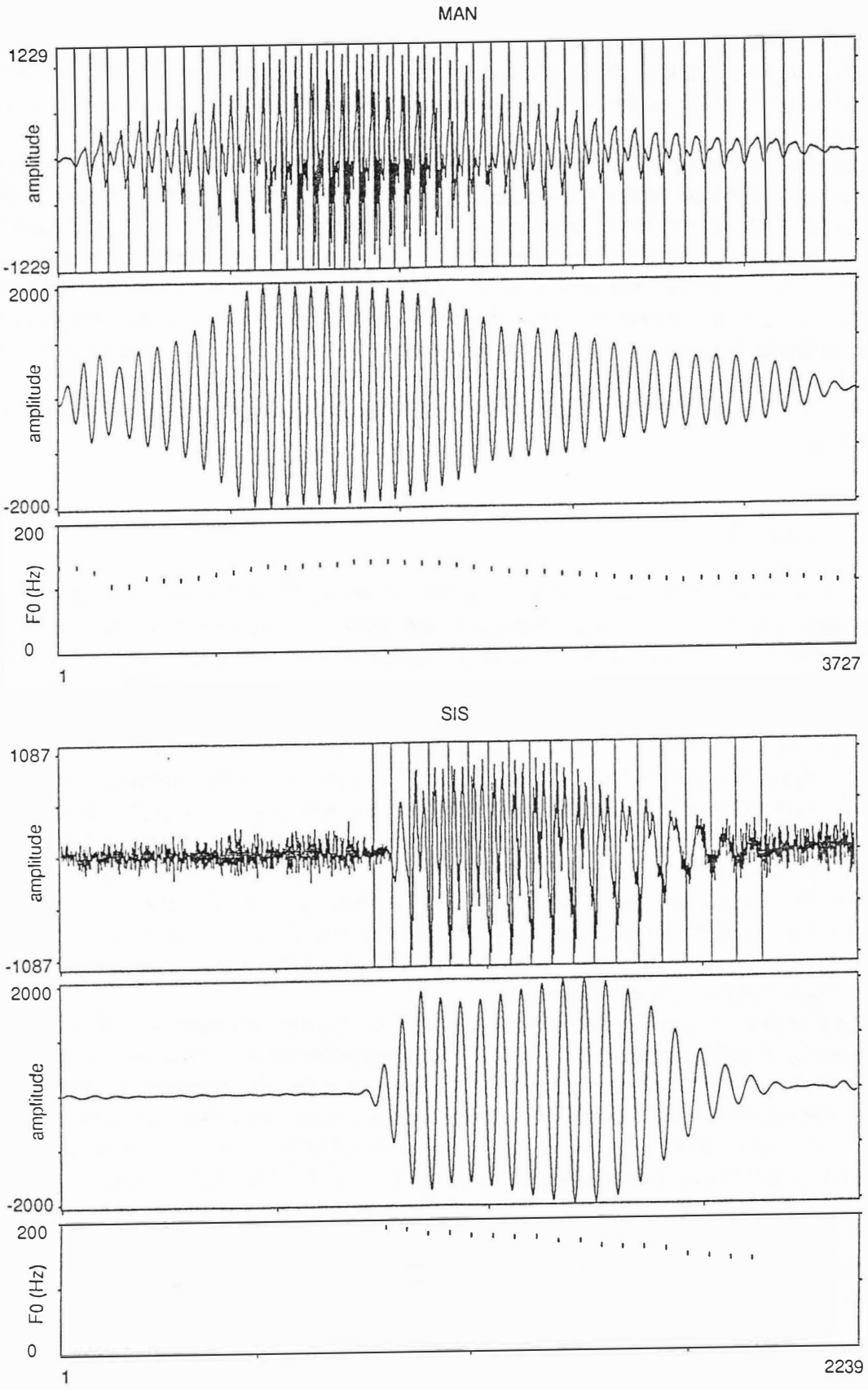

Figure 1. Two examples of the output of the new algorithm for the words "man" and "sis" uttered by a man to illustrate the positions at which the pitch markers are placed. The upper picture shows the speech signal with pitch markers, the middle picture shows the filtered output of the new algorithm and the lower picture shows the pitch contour.

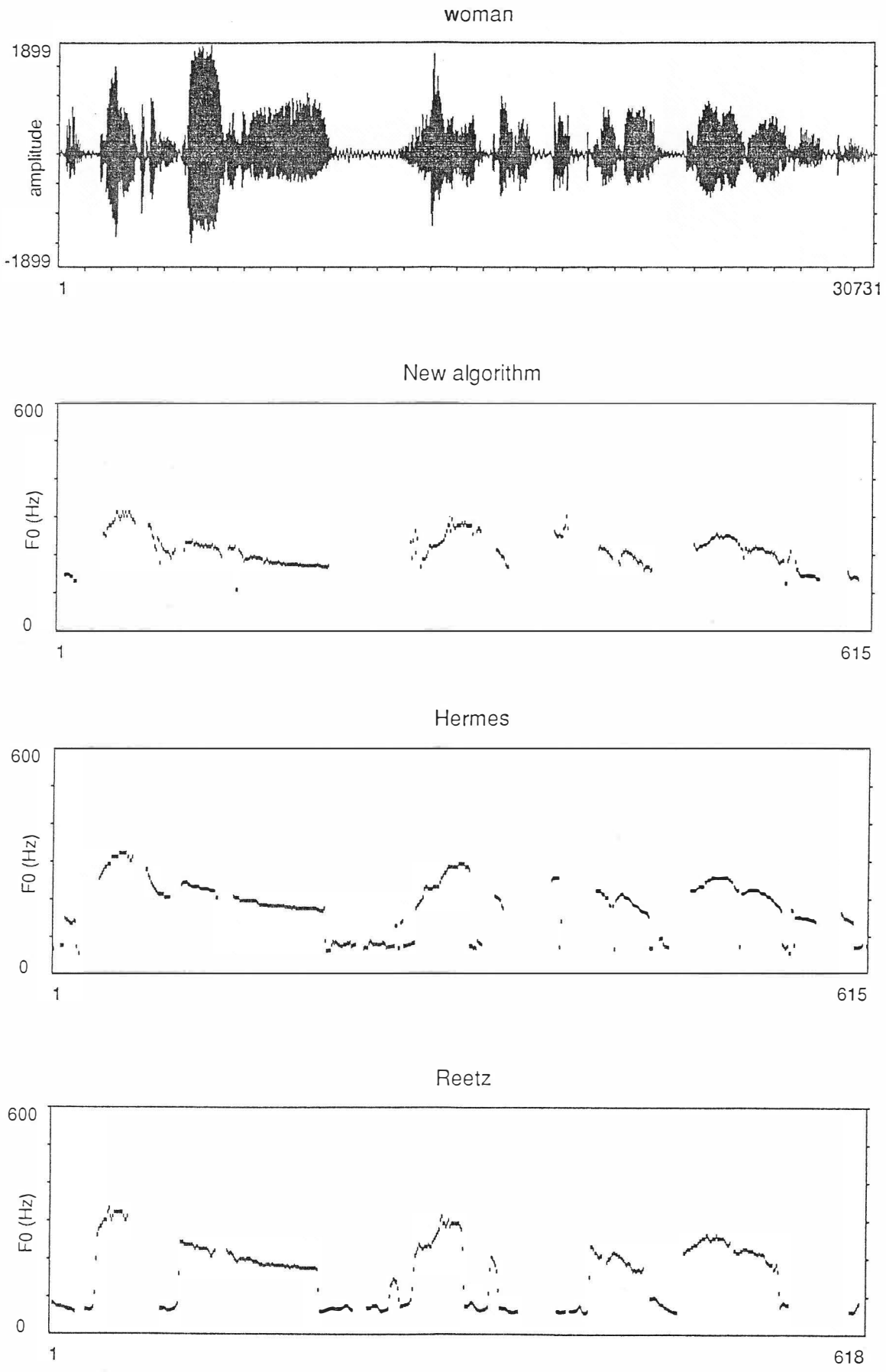

Figure 2. A 'field recording' of a female saying "Je kunt op deze manier leuke stukjes natuur ontdekken". The new algorithm is compared with the algorithms of Hermes (1987) and Reetz (1989).

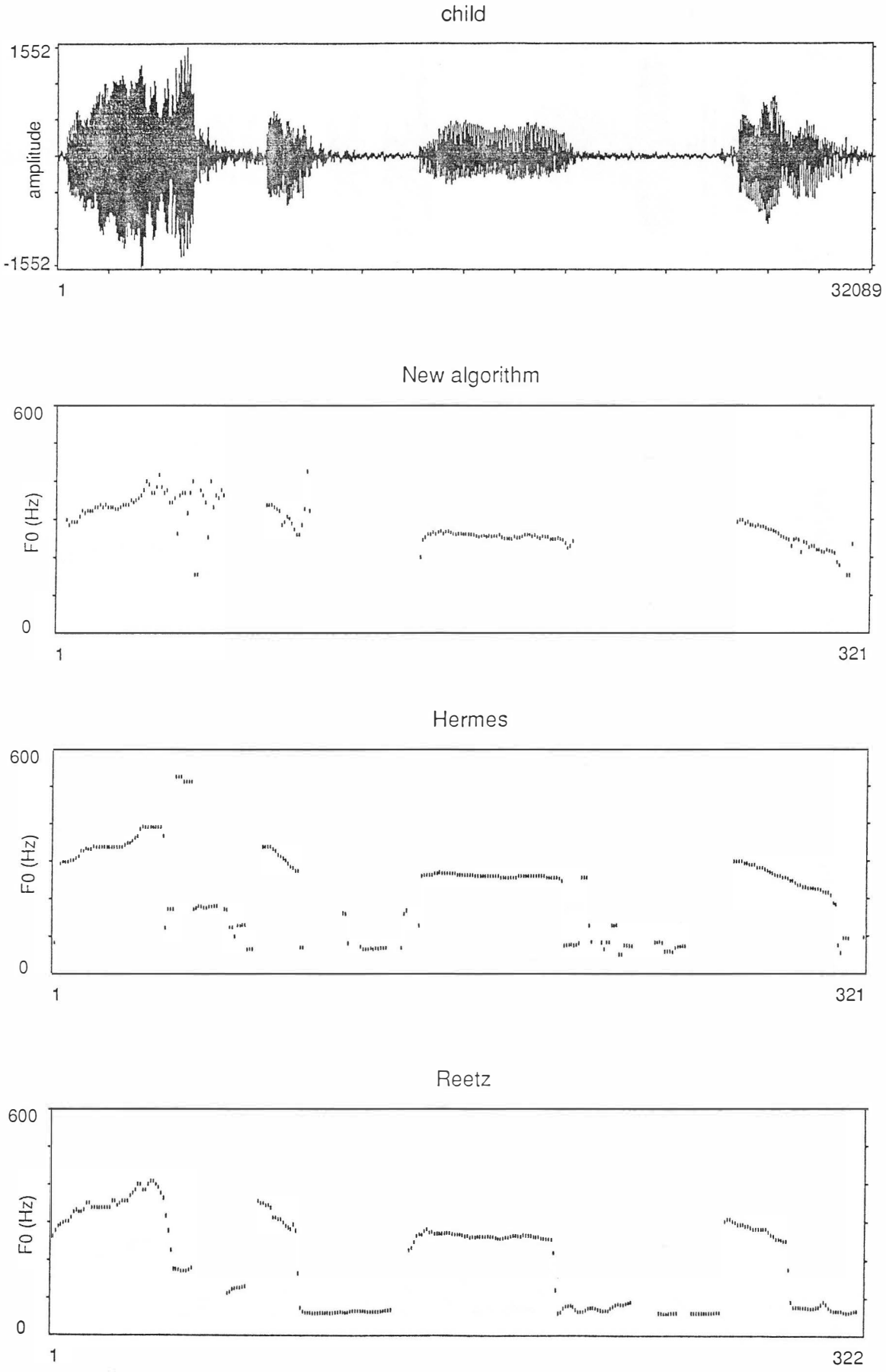

Figure 3. A 'field recording' of a 3-year-old child saying the word "brandweerauto". The performance of the new algorithm is compared with the algorithms of Hermes (1987) and Reetz (1989).

in the speech signal which results in a train of positive and negative peaks of varying amplitude. Subsequently a long series of tests are performed to eliminate those peaks that do not meet certain empirical criteria. The second pitch detector designed by Hermes (1988) estimates the pitch in overlapping frames of 40 ms by subharmonic summation.
Figure 2 shows the performance of the three algorithms for noisy female speech (a 'field recording'). The pitch contour provided by the new algorithm was checked by visual inspection of the pitch markers in the speech signal. Two markers were missing and a few markers were out of place due to incomplete attenuation of second harmonics. A first glance at the pitch contours provided by the algorithm from both Hermes and Reetz shows that a considerable part of these contours contain measurements of low frequency noise. Another thing to notice is the micro intonation given by the new algorithm which is due to the exact measurements of distances between successive pitch markers. The other two algorithms give only estimates of the local pitch through an averaging mechanism. In the third place parts of the contour are missing in the pitch algorithm from Reetz. In figure 3 the algorithms are compared for the noisy speech of a child (a 'field recording'). Visual inspection of the pitch markers provided by the new algorithm showed that one marker was missing and that a few markers were out of place.

5. CONCLUSIONS

In this paper an algorithm is presented that provides pitch markers in a speech signal at positions that coincide with the closure of the vocal chords, based on a method used by Dologlou and Carayannis (1989). No particular theory of human pitch perception is the root of this algorithm; the fundamental frequency is simply filtered out of the speech signal on the basis of a robust global estimate of the pitch contour provided by an autocorrelation analysis. Since the pitch is indicated by period markers, this method can be used for pitch synchronous applications and to measure pitch contours with high fidelity by visual inspection (and adjustment if necessary) of the markers. Extraction of the fundamental frequency through filtering implies that the method is not suitable for certain bandlimited signals such as telephone speech.
It is assumed that pitch movements are smooth and do not vary excessively within a 100 ms segment of speech (Remember that the bandwidth of the filter is half of the estimated fundamental frequency). For this reason pitch jumps (creaky voice) occurring in the speech signal are not properly processed. Although the algorithm was not tested extensively, it performed excellently in supplying pitch markers for sentences (high quality recordings) that were manipulated using the PSOLA-technique. It also compared favourably with two other pitch extractors in the analysis of speech recordings that were corrupted by noise.
A final remark concerns the calculation time required for the new method. No attempt was made to efficiently implement the algorithm, but it should be clear that the computational requirements are rather high because of the two-step analysis. However, if this point is of minor importance, this new pitch extractor appears to be a valuable addition to the existing ones.

6. REFERENCES

Atal, B.S. & Rabiner, L.R. (1976). "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, 201-212.

Charpentier, F. & Moulines, E. (1989). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Proc. Eurospeech '89, Paris Vol. 2, 13-19.

Crochiere, R.E. & Rabiner, L.R. (1983). Multirate digital signal processing. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Dologlou, I. & Carayannis, G. (1989). "Pitch detection based on zero-phase filtering", Speech Communication 8, 309-318.

Hamon, C., Moulines, E. & Charpentier, F. (1989). "A diphone synthesis system based on time-domain prosodic modifications of speech", Proc. ICASSP 1989, Glasgow Vol. 1, 238-241.

Hermes, D.J. (1987). "Measurement of pitch by subharmonic summation", J. Acoust. Soc. Am. 83, 257-264.

Laan, G.P.M. (1990). "Het belang van de spectrale kwaliteit van klinkers voor de verstaanbaarheid van zinnen", unpublished MA thesis, University of Amsterdam.

Rabiner, L.R. (1977). "On the use of autocorrelation analysis for pitch detection", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, 24-33.

Rabiner, L.R. & Schafer, R.W. (1978). Digital processing of speech signals. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Rabiner, L.R., Cheng, M.J., Rosenberg, A.E. & McGonegal, C.A. (1976). "A comparative performance study of several pitch detection algorithms", IEEE Trans. Acoust., Speech, Signal processing, Vol. ASSP-24, 399-418.

Reetz, H. (1989). "A fast expert program for pitch extraction", Proc. Eurospeech '89, Paris, Vol. 1, 476-479.