MEASURING PITCH PATTERNS WITHIN THE SCOPE OF EARLY SPEECH DEVELOPMENT

Els A. den Os and Florien J. Koopmans-van Beinum

ABSTRACT

This paper reports on a pilot experiment concerning the relation between pitch pattern transcription by hand and ear, and F0-measurements by means of a recently developed pitch extraction program (Reetz, 1989). The pilot study is part of a larger project on the development of pitch and temporal structure starting from the early lingual period.
Speech utterances of one child at the age of 1;3 and at 2;2 were analysed by the pitch extraction program. The resulting data were classified in terms of F0-contours and compared with the perceptually based pitch transcription of the utterances done by four experienced listeners. From this comparison a working strategy is developed to be applied in the continuation of this research project.

1. INTRODUCTION

The aim of the research project is to investigate the relationship between the development of pitch patterns and temporal structures on the one hand, and the acquisition of linguistic structures on the other hand. One possible hypothesis in this is, that 'new' syntactic structures might be expected on 'old' prosodic patterns, and 'new' prosodic patterns on 'old' syntactic structures, which is called the trade-off-hypothesis. An other, more cautious hypothesis assumes that the child initially produces stereotype pitch patterns of which the functionality, as present in adult speech communication, has to be learned gradually.

However, before we are able to test any hypothesis, we have to perform an exploring investigation on methods to be used, especially for the description of the development of pitch patterns.

Studying the development of intonation into communicative pitch patterns involves at least two main problems: a technical one and a perceptual one. The technical problem results from the difficulty to apply to high-pitched voices the standard pitch extraction programs, which are developed normally for male voices. Artifactal octave jumps and missing or evidently wrong F0-frequencies have to be detected and corrected by hand, a too time-consuming job within a large research project.

The other problem has to do with the well-known difficulty for listeners to reliably indicate the perceived (adult) intonation contours (cf. 't Hart, 1981). This problem, however, is even more serious in transcribing high-pitched voices, often making it necessary to work with a number of trained transcribers (cf. Koopmans-van Beinum and Van der Stelt, 1986). Nevertheless we need the judgement of the listener within our project, since after all it is the human listener in the environment of the child who defines the communicative content of the produced pitch patterns in development.

On the one hand we intended to rely for our intonation data as much as possible on acoustic measurements, but on the other hand we had to know how the F0-contours had to be interpreted in relation to the perceived communicative pitch patterns. So we needed solutions for both methodological problems, before reliable analyses of the prosodic development in relation to linguistic development could take place. On the basis of the recordings of one child we developed our method, as described below.

2. SPEECH MATERIAL

2.1. Recordings

During two years video- and audio-recordings were made monthly of four first-born children, two boys and two girls, starting from the moment the children were said to begin talking, according to the opinion of the parents. It happened that all children were 1;2 at the start of the recordings.The video- and audio-recordings were made at the homes of the children in a natural play situation, usually with the mother or sometimes with the father, and lasted about one hour each. The video-recordings were made with a Blaupunkt CR 1500 Videomovie, the audio-recordings with a Marantz CP-430 cassette recorder and two Sennheiser MD 21 microphones.

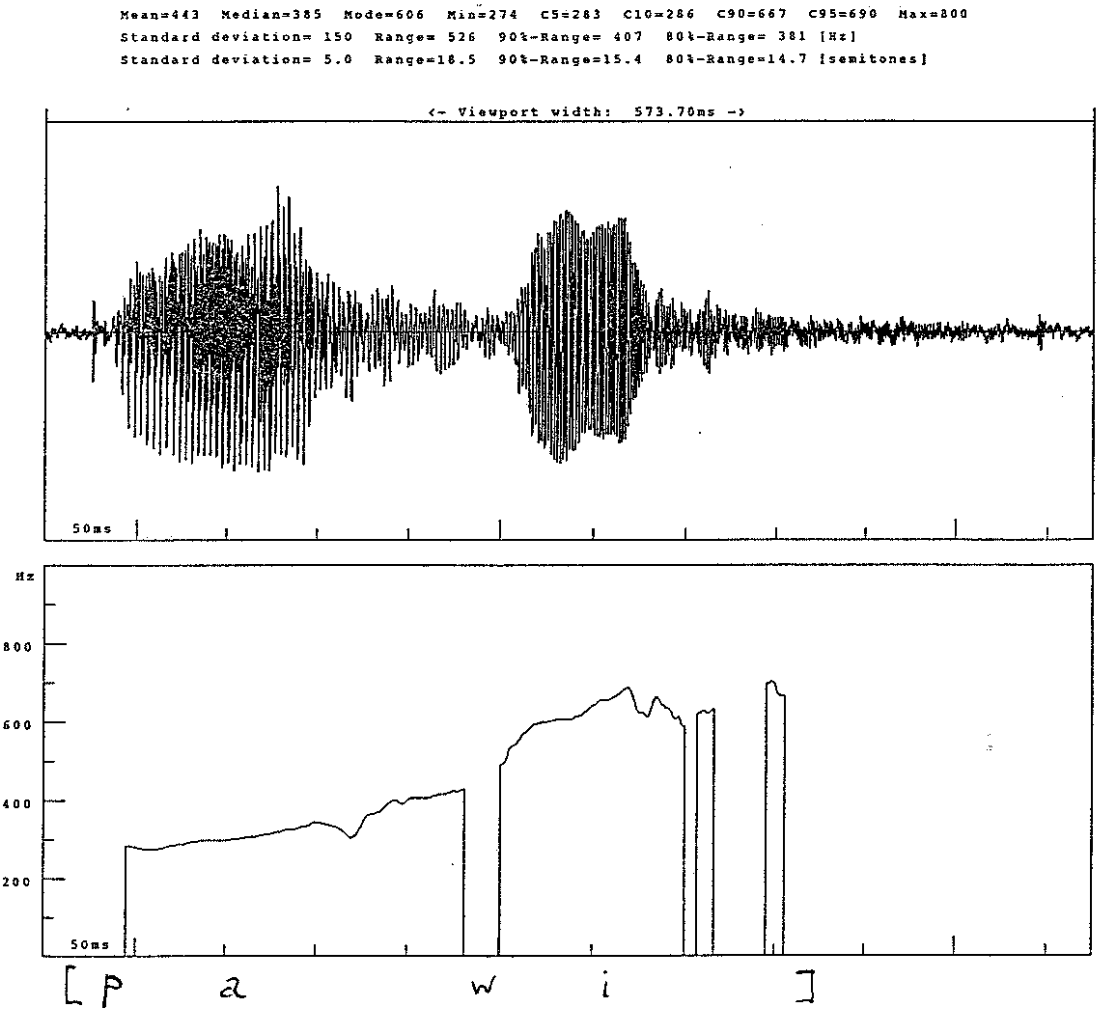
For this pilot study we used the speech material of one child, a boy DH, who at the age of 1;8 started to use two- and more-word utterances. The child started to use more-word utterances at the same time he started using verbs. The number of self-imitations in his utterances was rather low. To prevent that our selected methods would only suit the speech productions at a special age, we chose for this pilot one early (age 1;3;14) and one later recording (age 2;2;1). A continuous part of 20 minutes in the middle of each of the recordings was transcribed broadly with the help of the visual information of the video-recordings. The first 100 utterances were used for further analysis. Since not all utterances were acoustically appropriate for measuring because of background noises, we were left with 80 utterances of the early recording and 58 of the later one. The fact that especially in the later recording so much (more than 40%) could not be used for acoustic measurements is a consequence of the naturalistic recording setting and the growing age of the child. Since he displayed a great activity, playing with his toys and meanwhile communicating with his mother, and since at this stage he used mainly more-word utterances, his utterances often coincided with surrounding noises and reactions of his mother.

2.2. Acoustic intonation measurements

Although we knew that F0-measurements on the high-pitched voices of (very young) children might cause some problems, we preferred for our future research acoustic measurements since we believed this to save time in the long run providing us with more objective data, as compared to a more subjective, perceptual, evaluation of intonation done by a number of trained listeners.

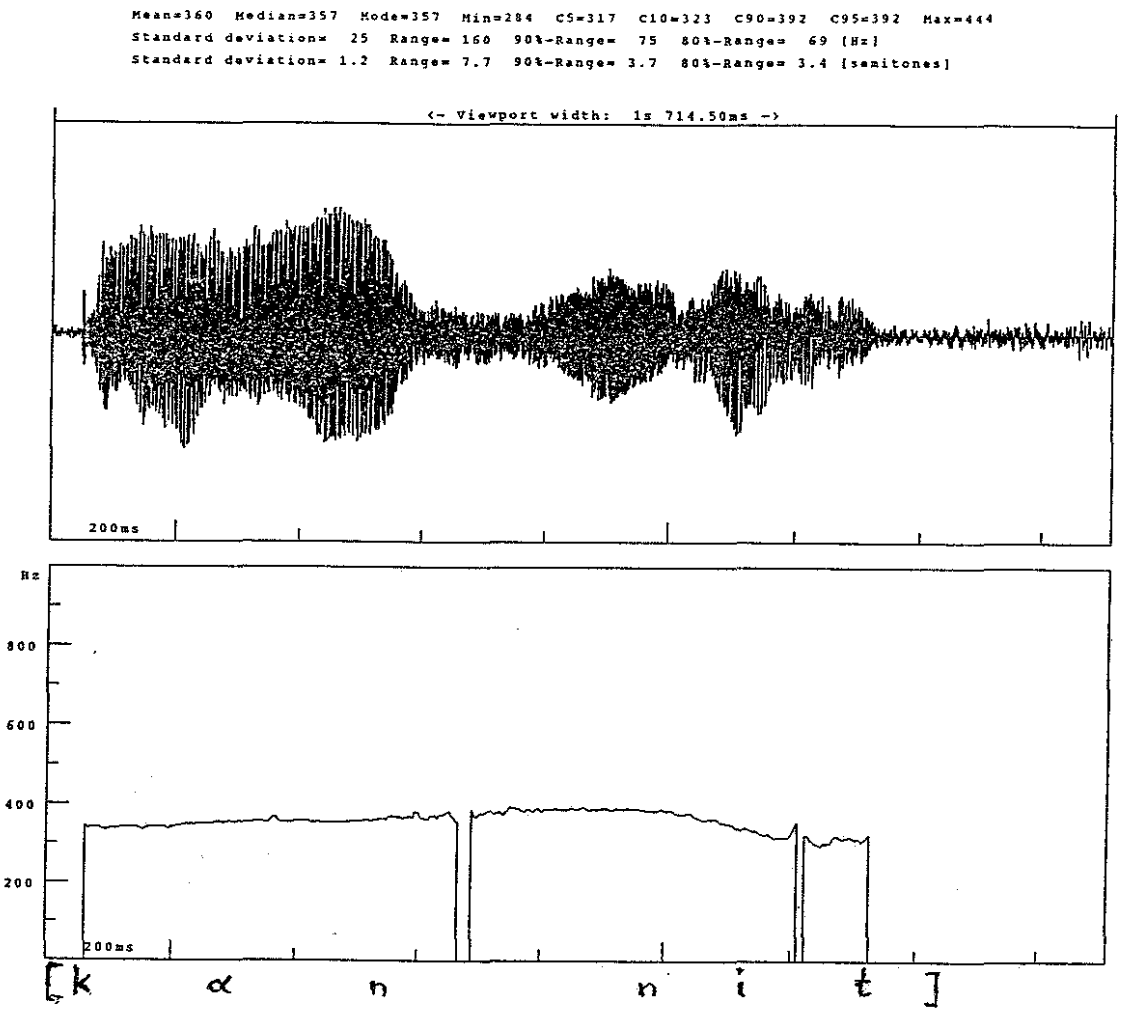
However, almost all common pitch extraction programs, frequency-based as well as time-based, turned out to be very time-consuming since many artifactal octave jumps and other obvious errors had to be corrected by hand. These errors are mainly caused by difficulties for the algorithms to handle the voiced-voiceless distinction, and by the sensitivity of the algorithms for background noise (see e.g. Tielen, this volume, p.49). The best working program for our purposes turned out to be the fast expert program for pitch extraction as developed by Reetz (1989), that in our opinion worked strikingly better than all other programs we tried to use. This algorithm operating in the time domain is said to be "resistant against (non-periodic) noise, and detects pitch reliable in a range between 50 Hz and 1000 Hz without parameter adjustment or voice/voiceless pre-segmenting" (p. 476). It combines information from the inspection of the speech waveform with knowledge about the glottal source signal. As described in Reetz (1989, p.476) the algorithm consists of four major steps: 1) Data reduction: converting the speech signal to prominent peaks. 2) Logical filtering: eliminating noise, amplitude and distance variations of the peaks. 3) Chaining: searching for an optimal pitch track through all peaks. 4) Post processing: eliminating too high, too low, or too short segments.

The program works left-to-right and right-to-left over a certain amount of data, delivering pitch values not synchronously with incoming speech samples. However, it works very fast, actually less than realtime on a Vax 750, and is in our experience quite promising since relatively very few corrections have to be made compared to other pitch extraction programs. In only a few cases it was necessary to increase the lower boundary since otherwise low frequency noise sources could be detected as a very low F0. Like most other pitch extraction programs this program provides, apart from a clear display of the waveform and the F0-contour, also mean F0-values and standard deviations over (part of) the whole utterance, and data on the range of the measured F0-values (see Fig. 1 and Fig. 2).



```
Mean=443  Median=385  Mode=606  Min=274  C5=283  C10=286  C90=667  C95=690  Max=800
Standard deviation= 150  Range= 526  90%-Range= 407  80%-Range= 381 [Hz]
Standard deviation= 5.0  Range=18.5  90%-Range=15.4  80%-Range=14.7 [semitones]
```

Fig. 1. Example of waveform and pitch contour of an utterance from the early recording, provided by the pitch extraction algorithm of Reetz (1989).



```
Mean=360  Median=357  Mode=357  Min=284  C5=317  C10=323  C90=392  C95=392  Max=444
Standard deviation=  25  Range= 160  90%-Range=  75  80%-Range=  69 [Hz]
Standard deviation= 1.2  Range= 7.7  90%-Range= 3.7  80%-Range= 3.4 [semitones]
```

Fig. 2. Example of waveform and pitch contour of an utterance from the later recording, provided by the pitch extraction algorithm of Reetz (1989).

3. TRANSCRIPTION EXPERIMENT ON INTONATION PATTERNS

Since it is our intention for the whole research project to base our description of the development of intonation on acoustic measurements, we need to know whether our transcription of the measured F0-contours is in agreement with the communicative pitch pattern as perceived by the listener. If the listeners, c.q. the care-takers of the child, are not able to discriminate small F0-movements and therefore will not use them in their communication with the child, it does not make sense to differentiate between those contours in the transcription of our measurements. To investigate this we concentrated on the interpretation of the contour of the final part of the utterances, defined as consisting of a lexical head (noun, verb, adjective, adverb) together with the non-lexical items going with it (Menyuk, 1972; Flax, 1986). This part is believed to provide the most important intonation pattern in communication (e.g. Lieberman et al., 1985), defining the reaction of the care-taker. Based on the achieved displays of the F0-contours the experimenter classified the contour as *rising*, *falling*, or *level*. An item was called *level*, if the standard deviation within the F0-contour was smaller than or equal to one semitone (Allen, 1983). If the F0 increased in the second half of the item and the standard deviation within the whole F0-contour of the item was more than one semitone, the item was called *rising*, and if the F0 decreased in the second half of the item and the standard deviation again was more than one semitone, the item was called *falling*.

Subsequently, four experienced female listeners, accustomed to listening to infants' and children's voices, were asked to indicate the perceived final F0-contours by using the same three possibilities as mentioned above and by interpreting the presented items as if spoken within a communicative setting. If they felt these three possibilities to be too limited to describe the perceived pattern, the listeners were allowed to use *rising-falling* and *falling-rising* as well.

As for the early recording this meant that the listeners were presented with the complete one-word utterances, whereas for the later recording only the last part was presented if the whole utterance was more than one word long.

4. RESULTS

4.1. Utterances from the early recording

From the recording made at the age of 1;3 all 80 items to be used for acoustic measurements were one-word utterances. Sample frequency was 20 kHz. The mean F0 was 372 Hz (s.d. 113 Hz) over all items together. Averaging within each item was done in steps of 10 msec. Based on the criterion mentioned above (Allen, 1983), the measured F0-contours were classified as:

level	23	(29%)
falling	20	(25%)
rising	37	(46%)

These classifications of the measured F0-contours were compared with the judgements in the listening test. For this purpose the *rising-falling* judgements, if used by the listeners, were put on a par with the *falling* ones, whereas *falling-rising* judgements did not occur at all.

The results of this comparison were as follows:

- for 35% of the utterances the judgements of all four listeners were in agreement with the measured F0-contours as classified by the experimenter;
- for 60% of the utterances the judgements of at least three of the four listeners were in agreement with the measured F0-contours as classified by the experimenter.

Two systematic deviations were found, when listeners' judgements were compared with measurement classifications, leading to the necessity to adjust our classification criteria.

Firstly, it turned out that in the case of monosyllabic utterances, *risings* were heard as *level*, unless the standard deviation over the whole utterance exceeded three semitones. This phenomenon presented itself only for *risings*. If the interpretations were corrected on the basis of this criterion, the results of our comparison were as follows:

- for 41% of the utterances the judgements of all four listeners were in agreement with the measured F0-contours as classified by the experimenter;
- for 70% of the utterances the judgements of at least three of the four listeners were in agreement with the measured F0-contours as classified by the experimenter.

So, subsequently, for the whole research project the three-s.d. criterion will be applied to the interpretations of F0-contours for all monosyllabic utterances.

The second systematic deviation consisted in the incapability of the listeners to differentiate between *falling* and *level* contours classified according to the criterion of Allen (1983): it often happened that measured *fallings* were judged to be *level*, whereas measured *level* contours were frequently indicated as *fallings*. So it might be better to differentiate only between *rising* and *non-rising* patterns in classifying F0-contours. From the communicative point of view this is quite tenable, since *level* and *falling* utterances will not result in different reactions in communication. Flax (1986) also concludes in a pilot study on intonation judgement of children's speech, that a classification of pitch patterns in *rising*, *level*, and *falling* is too much in detail.

If we apply the rising vs. non-rising criterion to our comparison we find:

- for 68% of the utterances the judgements of all four listeners to be in agreement with the measured F0-contours as interpreted by the experimenter;
- for 85% of the utterances the judgements of three of the four listeners to be in agreement with the measured F0-contours as interpreted by the experimenter.

These values in our opinion are quite acceptable (see e.g. 't Hart, 1981) for the purpose we need them for, taking into consideration the early linguistic stage of the child in this recording. The remaining 15% did not show any systematicity in their deviation, apart from the fact that in most of these cases only two out of the four listeners agreed.

4.2. Utterances from the later recording

From the recording made at the age of 2;2 no monosyllabic utterances occurred within the total number of 58 utterances to be used for acoustic measurements. Here the mean F0 over the final parts was 316 Hz (s.d. 39 Hz). Differentiating in *rising*, *falling*, and *level*, basing ourselves on the original criterion of 1 s.d. as mentioned above, resulted in the following classification of the measured F0-contours:

level	11	(19%)
falling	36	(62%)
rising	11	(19%)

Again the classifications of the measured F0-contours were compared with the judgements in the listening test. Since no monosyllabic utterances occurred in this recording, the three-s.d. criterion was not applied here. The results of the comparison were as follows:

- for 55% of the utterances the judgements of all four listeners were in agreement with the measured F0-contours as classified by the experimenter;
- for 78% of the utterances the judgements of at least three of the four listeners were in agreement with the measured F0-contours as classified by the experimenter.

Here again we found a systematic confusion by the listeners of the *level* and *falling* contours as classified by the experimenter. Therefore, we decided to use only the distinction *rising* and *non-rising* here as well.

After this adjustment the results we found:

- for 76% of the utterances the judgements of all four listeners to be in agreement with the measured F0-contours as classified by the experimenter;
- for 93% of the utterances the judgements of at least three of the four listeners to be in agreement with the measured F0-contours as classified by the experimenter.

As could be expected, it turns out that it is easier for listeners to make judgements in a more uniform way as the utterances become longer, more meaningful, and more adult-like.

5. CONCLUSION

This pilot study was intended to investigate the possibilities of measuring F0-contours in developmental speech in a quick and reliable way, showing to full advantage the perception of the trained listeners, when judging the intonation patterns of early-lingual, communicative utterances.

Since measured F0-contours have to be classified into linguistically relevant categories, it was thought sensible to find first of all a satisfying pitch extraction method in order to classify the resulting F0-contours, especially the final ones, and then subsequently relate these classifications to the perception results of listeners, who were asked to classify the perceived intonation patterns into the same categories.

The pitch extraction method of Reetz (1989) was found to be a fast and most satisfying program for high-pitched voices, providing clearly interpretable F0-contours.

The comparison of initially categorised F0-measurements with listener judgements required several steps of adjustment of the classification criteria, and a reduction of pitch pattern categories into rising and non-rising contours, but then performed satisfactory.

Since the method defined in this pilot study has been tested on an early recording as well as on a later one, it will in our opinion be acceptable to use this method in our whole research project on the development of prosody in the early-lingual period.

ACKNOWLEDGMENT

We wish to thank the direction of the Max Planck Institute for Psycholinguistics in Nijmegen, and Henning Reetz, for making it possible for us to use their speech analysis facilities and especially their pitch extraction program.

REFERENCES

Allen, G.D. (1983). "Some suprasegmental contours in French two-year-old children's speech". Phonetica 40, 269-292.

Flax, J.F. (1986). Functional intonation in the prelinguistic and the early linguistic child. Unpublished Ph. D., City University of New York.

't Hart, J. (1981). "Differential sensitivity to pitch distance, particularly in speech". J. Acoust. Soc. Am. 69, 811-821.

Koopmans-van Beinum, F.J. and Stelt, J.M. van der (1986). "Early stages in the development of speech movements". In: B. Lindblom & R. Zetterström (Eds.), Precursors of early speech. Wenner Gren International Symposium Series, vol. 44, Basingstoke, Hampshire, England; Macmillan Press, 37-50.

Liebermann, P., Katz, W., Jongman, A., Zimmerman, R., and Miller, M. (1985). "Measures of the sentence intonation of read and spontaneous speech in American English". J. Acoust. Soc. Am. 77, 649-657.

Menyuk, P. (1972). The development of speech. The Bobbs-Merrill Company, Inc., Indianapolis, New York.

Reetz, H. (1989). "A fast expert program for pitch extraction". In: J.P. Tubach and J.J. Mariani (Eds.), Proceedings of Eurospeech 89, European Conference on Speech Communication and Technology, Vol. 1, 476-479.

Tielen, M.T.J. (1989). "Fundamental frequency characteristics of middle aged men and women". Proceedings of the Institute of Phonetic Sciences Amsterdam 13, 49-58 (this volume).