A Perceptual Evaluation of Two V/U Detectors

N. van Rossum and A. Rietveld
Nijmegen, the Netherlands

1. Introduction

In many analog pitch meters, however different they may be, the pitch detection circuitry is controlled by a voiced/unvoiced (V/U)-detector. This is to say that pitch will only be determined in those segments which have been labelled 'voiced' in a previous stage.

Both parts of such pitch meters, the voiced/unvoiced decision system and the pitch detector itself, have one thing in common: the evaluation problem. In both cases it is difficult to find clearly operationable correlates of the features to be detected in the acoustical and perceptual domains.

In this contribution we will focus on the description and perceptual evaluation of two V/U-detectors which are parts of two analog pitch meters described elsewhere (van Rossum, 1982). We tried to find an answer to the following questions:

1. Is it possible to obtain reliable 'voiced/unvoiced' judgments from a panel of listeners?
2. Do judges agree equally well on the onset and offset of voiced segments?
3. Which of the two V/U detectors (to be described below) corresponds best with the decisions of the listeners?

2. Short Description of the two V/U-detectors

We tested two different detectors which are integral parts of two analog pitch processors developed in our laboratory. Fig. 1 gives a blockdiagram of these V/U detectors; their main characteristics will be summarized below.

a. A classical V/U detector, in which the energy in a low frequency band (20 Hz - 1 K Hz) is compared with a predetermined criterion. This detector is based on a principle already applied by Dudley (1939); it was found to be very reliable by Wiren and Stubbs (1956).

b. A V/U detector which measures the spectral balance in the speech signal. To this aim the energy difference in the bands 20 Hz - 1 kHz and 5 kHz - 14 kHz is determined, whereafter the result is compared with a criterion value. Voiced segments are assumed to have predominantly low frequency energy and voiceless segments high frequency energy.

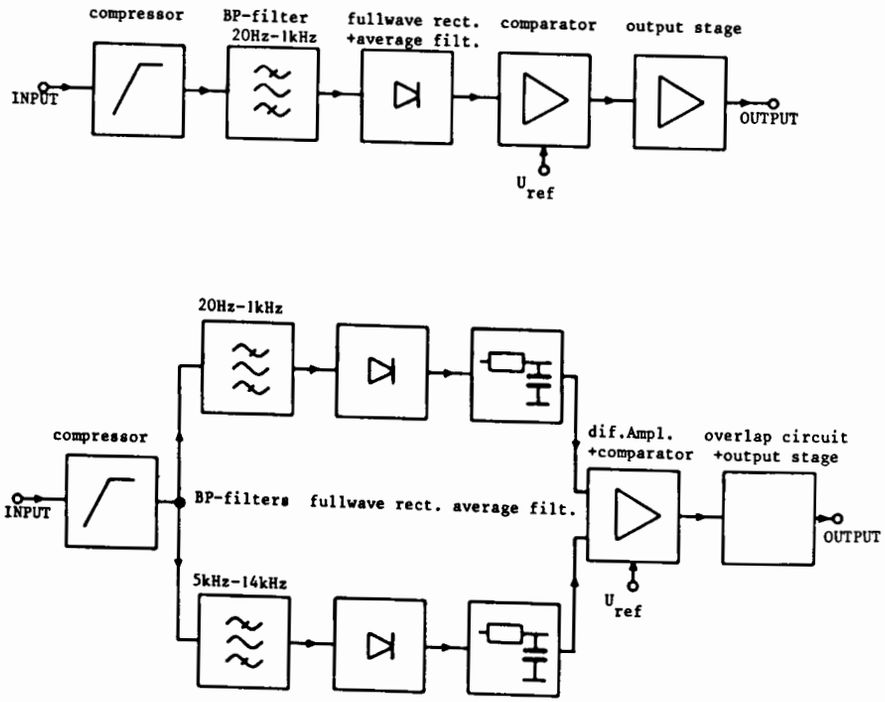

Figure 1. Blockdiagrams of a 'classical' V/U-detector (A) and of an energy difference V/U-detector (B).

This type of detector was originally designed by Knorr (1979) and has been adapted by us. The detector does not only differ from the previous one in the frequency bands which are used, but also in an 'overlap circuit' applied to the output of the detector. In order to avoid 'jittering' in the V/U-output, this circuit was integrated in the design; the resulting delay time amounts to about 10 ms, depending on the input waveform.

Knorr's evaluation of the detector was based on the comparison of the V/U-decisions and the waveform; the result of this evaluation was reported to be very satisfactory. In our opinion, however, it is rather difficult to determine whether a signal is semiperiodic - especially at the boundaries of voiced segments - and should consequently be labelled as 'voiced'.

For that reason we designed another evaluation test, a perceptual one.

3. The perceptual evaluation of V/U-detectors

The evaluation of V/U-detectors is notoriously difficult for many reasons; the main problem is the absence of a one-to-one relation in the three domains involved. Dolansky (1968) showed that the vibration of the vocal cords does not always result in a periodic signal, whereas Glave (1973) found that stochastic signals without a clear periodicity may have a 'tonal' quality and

will consequently be judged as voiced. Furthermore, the distinction tone/no tone appeared not to be a categorical one.

A supplementary problem resides in the fact that different evaluation procedures may lead to different results. The scores obtained in a perceptual evaluation of a V/U-detector will depend on the size of the signal units which are to be judged. Segments of phone size will be judged in a totally different manner than segments of say 30 ms. In the former case a complex of cues will be used (Slis and Cohen, 1969), in the latter case the judgments will be based on spectral features.

If a perceptual evaluation is realized by means of speech resynthesis in which voice-onsets and offsets are controlled by a V/U-detector, the two possible biases of the detector, voiced and unvoiced, will differently influence the acceptability of the synthesized speech signal.

An 'acoustic' evaluation is also rather hard to perform. Proper periodic signals hardly occur in real speech; at the end of voiced segments semiperiodic signals cannot easily be distinguished from noise.

We chose a perceptual criterion in the evaluation of the V/U-detectors involved, one reason being the difficulties which can be expected in an acoustic evaluation. Our experiment was designed in such a way that subjects had to judge short successive speech segments of 30 ms as 'voiced' or 'unvoiced'. We chose this perceptual scanning procedure because it simulates to a certain extent the functioning of the V/U-detectors; these detectors determine whether short speech segments (in fact indefinitively short) will be further processed by the pitch detector or not.

4. Procedure

Before the real trials started, the subjects were given a set of 'anchoring' trials in which short segments (30 ms) of clearly voiced or unvoiced speech were presented.

The experimental trials consisted of segments from six sentences each spoken by two speakers (one male, one female).

By means of a variable gate successive segments of 30 ms were presented over earphones to ten listeners (5 male, 5 female). By means of a thumbwheel switch the subjects could shift a segment by incremental steps of 10 ms. The subjects had to mark the transitions from 'voiced' to 'unvoiced' segments. The thumbwheel switch indicated the onset of the trapezoidal window, as is shown in Fig. 2.

In this way a perceptual scanning of the twelve sentences took place; the resulting data were the perceived onsets and offsets of the voiced segments in milliseconds.

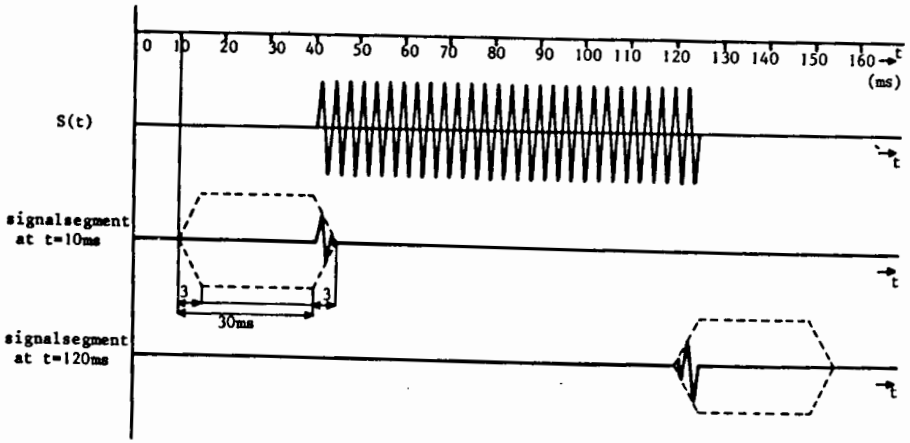

Figure 2. Position of the trapezoidal window and the perceived onset and offset.

5. Results

The perceptual scanning resulted in 78 segments which were labelled 'voiced' by at least two of the ten listeners. For further processing we only used the registered voice onsets and offsets of those segments which were judged 'voiced' by at least 7 of the ten listeners (62 segments).

Subjects were found to be reliable in their judgements. Ebel's reliability coefficient was .99 for both onset and offset times. Therefore, we feel justified to make use of the mean scores in the subsequent analyses.

In spite of the high reliability, an F-test showed differences in agreement between the perceived onset and offset times: $F = 18.77$ $df1/2=9$, $p < 0.001$. It appeared that subjects agreed more on the onset of voiced segments than on the offset; this difference is probably due to the often observed asymmetry of the waveform of voiced segments.

As for the agreement between the decisions of the listeners and both detectors, we performed separate analyses for onset and offset times. In Table I we present the mean onset and offset times perceived by the listeners and those found by detector 1 and detector 2, respectively.

A Multiple Range Test (significance level: 0.05) showed significant differences between the onset times determined by the detectors and the perceived onset times; the difference between the detectors was not significant.

Table I. Onset and offset times of voiced segments in milliseconds; mean values of the listener's judgements arbitrarily set to zero

	listeners	Detector 1	Detector 2
onset	0	33	32
offset	0	2	6

The same test was applied to the offset times. The only significant difference was that between detector 2 and the listeners, a difference of not more than 6 milliseconds.

6. Discussion and Conclusion

The difference between the performance of the detectors and the decisions of the listeners on the onset of voiced segments can be explained by the method we used in the perception experiment. The registered onset times of the listeners equal the opening time of the variable gate. If a subject labels a segment 'voiced' as soon as the last part of the window is voiced, the registered onset time will be 30 ms ahead of the 'real' onset. An interval of about thirty milliseconds happens to be the difference we found between the decisions of the listeners and those of the detectors. As in most cases a close agreement was observed between the onset of semiperiodicity and 'voiced' labels of the detectors, we may conclude that the observed difference is for a great part due to the window we used in the experiment.

As for the offset times, the fact that the 'voiceless' decisions of detector 2 were significantly later than those of the subjects can be explained by the overlap circuit which is part of the system. This circuit appeared to be rather important as without it (detector 1) much more and longer jitterings - about 100% - were found. If we take into account the above mentioned effects, we may conclude that the two V/U detectors performed equally well and in close agreement with the judgments of the listeners. This finding should not obscure the existing differences between both detectors. In particular the overall-amplitude of the signal has a noticeable influence on the functioning of the detector which only operates on the LP-frequency band (detector 1) and much less influence on the performance of detector 2. For that reason, the latter should be preferred to the former.

As is well known, voiced/unvoiced decisions have a strong influence on the quality of synthesized speech. It is not yet clear whether parameter estimation in which our V/U detectors are used, will lead to acceptable resynthesized speech. Experiments in that direction are planned.

References

Dolansky, L. and Tjernlund, P. (1968). On certain irregularities of voiced speech waveforms. *IEEE Trans. Audio and Electroacoust., Vol. AU-16*, 51-56.

Dudley, H. (1939). The Automatic Synthesis of Speech. *Bell Teleph. Syst. Techn. Publ. Monograph B-1169.*

Glave, R.D. (1973). *Untersuchungen zur Tonhohenwahrnehmung Stochastischer Schallsignale.* Hamburg: Helmut Buske Verlag.

Knorr, S.G. (1979). Reliable Voiced/Unvoiced Decision. *IEEE Trans. on Acoustics Speech and Signal Processing, Vol. ASSP-27*, 3, 263-267.

Rossum, N. van and Boves, L. (1978). *An analog pitch-period extractor.* IFN-proceedings, **2**, 1-17.

Rossum, N. van (1982). Technische beschrijving van de F_0-processor. Interne Publicatie IFN.

Slis, J.H. and Cohen, A. (1969). On the complex regulating the voiced-voiceless distinction. Part I. Part II. *Language and Speech* **12**, 80-102, 137-155.

Wiren, J. and Stubbs, H.L. (1956). Electronic Selection System for Phoneme Classification. *JASA* **28** (6), 1082-1091.