F0 ANALYSIS PROBLEMS

Sidney Wood
Department of Linguistics and Phonetics
University of Lund

ABSTRACT
This paper reports an investigation of the sensitivity of speech analysis programmes to voice quality and signal quality, with a view to optimizing parameter settings for successful F_0 analysis. Results were more successful with data frame lengths of 10 ms than 5 ms and with better, less breathy, voices. In the speech samples used here, voice quality seemed to be more important than signal quality.

INTRODUCTION AND PROCEDURES

It is well known that F_0 analysis by ILS and similar programmes sometimes yields incongruous results with regard to phonetic properties like voicing and intonation. For example, they may fail to recognize the presence or absence of voice tone, on both vowels and consonants, and they may propose a high soaring tone on voiceless sibilants and plosive bursts. Anomalous analyses can always be cleaned up by editing the results, but this will obviously lead to circularity in some experimental contexts. It is particularly distressing if the analysis is required for a speech recognition device that is expected to identify segmental voicing/voicelessness or interpret an intonation pattern.

Obviously, phenomena like creak and vocal fry are readily identifiable discontinuities in the waveform that are perfectly audible disruptions to the otherwise melodic vibration of the voice tone. Cases like this are dealt with at length by Huber elsewhere in this volume. The puzzling cases are more subtle and examination of an expanded segment of the waveform does not always reveal just why the programme has overlooked the vocal vibrations during a vowel or has settled for harmonic vibration during the aperiodic hiss of a voiceless consonant. Sibilants and voiceless stops are particularly tricky.

Some voices seem to be more successfully analysed than others, suggesting that some personal factor is involved. One such

factor might be the degree of breathiness of the voice, which superimposes an aperiodic component and weakens the higher partials. Studio recordings from two speakers were therefore compared, both adult male native speakers of Bulgarian who read the same material, the one with a relaxed and non-breathy voice and the other very breathy.

One might also expect the quality of the recording to have some influence on the result, depending on the amount of distortion and on the signal/noise ratio. The presence of mains hum may also confuse F_0 analysis. A studio recording of originally good quality, of a trained professional speaker of Swedish, was degraded by being copied in several generations on primitive equipment in order to obtain a bad copy for comparison with the studio recordings of the two Bulgarian speakers.

As far as the programmes themselves are concerned, the speech samples were analysed by two methods for comparison: the ILS routines API (a modified cepstral processing technique) and SIF (the SIFT algorithm).

Finally, there are many programme parameters that can be set, and it need not be the case that the default settings are the optimum. The digitalized waveform is quantized into *data frames* of optional length, the longer the frame the poorer the temporal and spectral resolution but also the greater the chance that it will include at least one fundamental period. The default frame length, 64 data points, is optimized for input and output operations, not analysis. Frame lengths of 5 ms and 10 ms were used here. The LPC *analysis window* , of recommended length 15-35 ms, starts simultaneously with the data frame. It was set at 20 ms for this investigation. The analysis window is straddled by the *excitation buffer* that was kept at its default length of 32 ms; this buffer must be longer than the analysis window, around which it is shifted as the programme determines the periodicity of the current data frame. The periodicity decision depends on an excitation index, computed as a by-product of the cepstral analysis, exceeding a preset *excitation threshold*. This threshold (default 0) can be set by the user. Raising it sufficiently will purge the sequence of voiceless segments spuriously analysed as voice.

RESULTS AND DISCUSSION

With a *data frame* size of 50 ms, 11 out of 12 analysed sentences exhibited unexpected deviations in the F_0 curve. Lengthening the frame to 100 ms worsened one analysis, left 3

unchanged, somewhat improved 5 and produced an ideal result in 3. Lengthening the data frame thus improves F_0 analysis.

The choice of *analysis method* was inconclusive. Sometimes API gave a better result and sometimes SIF, apparently unrelated to the other factors investigated.

The results of the *poor signal quality* samples were only just slightly worse than those of the good studio recording of the relaxed non-breathy voice, and definitely better than those of the breathy voice. The speaker of the poor recording copy had a good, trained professional voice so that voice quality seems to be more important than signal quality for F_0 analysis.

The value of the calculated excitation index seemed to be slightly lower for longer data frames, which partly explains why lengthening the data frame improved the F_0 analysis. Raising the *excitiation threshold* certainly removed spurious voiced decisions, but often at the expense of losing some correctly voiced decisions. There may be an ideal setting for each individual voice, in which case the programme will need to be tuned to the speaker. Clearly, if the threshold has to be set differently for each sentence, we are back to an arbitrary and circular procedure again. One can certainly agree with the author of the ILS user notes for API, that analysis of the fundamental is an art and not science.

REFERENCES

Huber, D. (1988). Laryngealization as a boundary cue in read speech. P. 66 in this volume.

ILS. *The Interactive Laboratory System for Speech Analysis.*