Pilot study of F0 extraction errors made by Praat's autocorrelation algorithm

Aims

Experimental study of intonation requires reliable estimation of F0 contours extracted from audio samples. Most algorithms available through commonly used speech analysis software show good performance, but they are not error-free. Some level of human intervention is required in a number of cases. Correction of automatically extracted contours by humans is a time-consuming task that can be a bottleneck in the collection and study of large corpora, especially those recorded more naturalistic conditions. Common errors such as doubling or halving of F0 values can bias estimation of average F0 level or range that are important in applications such as voice comparison in forensic contexts. In this ongoing study, we aim to evaluate the performance of Praat's autocorrelation F0 extraction algorithm in order to assess what are the most common errors and how they may affect prosodic analysis. Praat was chosen because it has become the de facto software tool in speech analysis.

Methodology

Five females and five males provided samples of three speaking styles: an interview, read sentences taken from the interview and a list of words taken from the sentences. F0 contours for the audio files in each speaking style were extracted using the default parameters by means of Praat's "To Pitch" function. These contours were labelled the "raw" condition. Raw contours were then verified by a human analyst trained in the task of finding errors in the F0 estimates produced by the algorithm. The analyst looked for occurrences of abrupt change in F0, such as doublings and halvings. Suspect values were checked against the waveform and corrected when possible. When a suspect value was consistent with the waveform it was kept as is, but unvoiced when the analyst could not confidently identify periodicity in the waveform. Conversely, the analyst changed the voicing status of chunks deemed periodic but misidentified as unvoiced by the algorithm and a proper value was assigned. After the correction by the human analyst, the raw and corrected contours were compared. Each analysis frame was checked against two kinds of errors:

- false voicing: voiced frames present in the raw contour but unvoiced in the corrected contour;
- false unvoicing: unvoiced frames in the raw contour but voiced in the corrected contour.

When analysis frames were voiced in both contours, the following measure was calculated (F_r and F_c are respectively the values in the raw and corrected contours): $\log_2(F_r/F_c)$. Histograms of F0 values for each contour in the corpus were also generated to allow the visual inspection their distribution.

Results

Preliminary results show that for the interview style both false voicing and false unvoicing errors tend to occur at a mean rate of around 2.2 per second. Median differences in voiced frames tend to be well below 0.1 octave, with a maximum of half an octave for one male speaker. Halvings and doublings that observed in the histograms were all identified by the human analyst as instances of false unvoicing. In the next phase of the study, we are going to look into the possible performance differences caused by the other speaking styles as well as estimating the impact of the errors on long-term estimates of F0 central tendency such as mean and median. There has been suggestions in the literature that changing the so-called floor and ceiling parameters of the algorithm can lead to less extraction errors (Hirst 2011). In order to verify these claims we are going to systematically vary the two parameters and try to establish the range of values that most approximate the human corrected contours.

Reference

Hirst, D. J. 2011. "The analysis by synthesis of speech melody: from data to models", Journal of Speech Sciences, 1 (1): 55-83.