A comparison between three commonly used methods for pitch extraction in speech

Sofia Strömbergsson

Pitch extraction (or pitch tracking) underlies all data-driven analysis of intonation in speech. Although large datasets require automatic procedures, there are known uncertainties involved in using existing methods, raising reliability as an obvious concern. Considering their well-spread use in analyses of conversational prosody, getF0 (Talkin, 1995), Praat (Boersma, 1993), and YIN (de Cheveigné & Kawahara, 2002) often emerge as the top three candidates for pitch analysis, with none of them being a gold standard. This investigation set out to bench mark all three against a "ground truth" reference, and thereby, to illuminate the consequences of selecting one method over the other two.

The PTDB-TUG speech corpus (Pirker et al., 2011), consisting of 4720 sentences recorded from 20 healthy speakers (10 male and 10 female) with both a 48 kHz microphone signal and a parallel high-pass filtered laryngograph signal was used as a reference. The difference between the f0 traces as generated by the three candidate trackers and the ground truth laryngograph signal available in the reference corpus was evaluated with regards to several metrics (in accordance with Babacan et al., 2013). These metrics reveal, for example, the proportion of frames with an incorrect voiced/unvoiced decision, and – in cases where a voiced decision is correct – the difference between a candidate f0 value and the actual f0 value.

The results show that, of the three pitch extraction methods, the Praat pitch tracker outperforms the other two on all measures, although to varying degrees across measures. For example, the proportion of incorrect voicing decision was 4.8% for Praat, as compared to 6.2% (getF0) and 33.5% (YIN). Regarding the proportion of frames with gross pitch errors (i.e. relative pitch error > 1 semitone), the Praat tracker's performance was 1.5%, as compared to 2.6% (for both getF0 and YIN).

Although the selection of pitch tracker may be guided also by other concerns (e.g. user-friendliness or processing requirements), the reliability of the f0 traces is, of course, central. Considering that automatic analysis of pitch will always involve some error, quantification of the extent and variation of this error across different methods is important. The present investigation illustrates an approach to standardize the evaluation of pitch extraction methods, and reveals that in the choice between Praat, getF0 and YIN, the Praat pitch tracker is favored from the perspective of reliability.

References

Babacan, O., Drugman, T., D'Alessandro, N. Henrich, N. & Dutoit, T. (2013). A comparative study of pitch extraction algorithms on a large variety of singing sounds. In *38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Vancouver, Canada.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences 17* (pp. 97-110). Amsterdam, The Netherlands.

de Cheveigné, A. & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustic Society of America 111*(4), pp. 1917-1930.

Pirker, G., Wohlmayr, M., Petrik, S. & Pernkopf, F. (2011). A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario. In *Proceedings of Interspeech 2011*. Florence, Italy.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn & K. K. Paliwal (Eds.) *Speech Coding and Synthesis*. Amsterdam, Elsevier Science, pp. 495-518.