Pitch Tracking

MUMT 621 Presentation 4. March 16, 2021. Sevag Hanssian, 260398537

Summary

Pitch is the perceptual correlate of frequency (Plack 2013). According to Plack, the American National Standards Institute's definition of pitch as "that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high" is too ambiguous. The terms "high" and "low" can also apply to loudness and intensity. He proposes an alternative definition, that pitch is the aspect of auditory sensation whose variation is associated with musical melodies. For a pure tone consisting of a single frequency, the pitch of the tone is directly related to its frequency. For a complex tone with multiple frequency components, the pitch is related to the fundamental, or lowest frequency. The fundamental frequency is also referred to as *f0*.

Pitch is essential to music perception in humans (McDermott and Hauser 2005). Relationships between pitches are more important than the absolute pitch – humans in experiments can recognize melodies when they are transposed in tempo or octave. Pitches separated by the special octave relationship have the same pitch chroma. Most known music in the world across cultures and eras consists of melodies formed from five to seven pitches within an octave range.

Moore (2013) states that for years there were two theories on human pitch perception: place and temporal. The place theory, or place coding, is related to the tonotopic organization of the inner ear: spectral analysis is done in the cochlea, and the resolved harmonics of a sound excite different parts of the basilar membrane (BM), causing with different characteristic frequency (CF) to fire. The activated center frequencies may then be compared to a pattern of harmonics (but this second postulate is under debate). In the temporal theory, the unresolved harmonics form a complex waveform in the BM, and firing neurons lock to the phase of the envelope of the complex waveform. More complete, newer models that try to account for all the available experimental data rely on both place and temporal analyses.

In computational pitch tracking, fundamental frequency (*f0*) and pitch are used interchangeably.

Computational pitch tracking has been studied for at least half a century (Noll 1967). A common family of approaches involves selecting candidates from a generating function, and performing pre- and post-processing to produce the pitch curve – among the candidate functions are cepstrum (Noll 1967), autocorrelation function (Dubnowski et al. 1976), average magnitude difference function (Ross et al. 1974), normalized cross-correlation function (RAPT and PRAAT; Talkin 2005; Boersma 1993), cumulative mean normalized difference function (YIN; Cheveigné and Kawahara 2002), and the normalized square difference function (McLeod and Wyvill 2005). Some forms of post-processing of the pitch candidate functions involves peak picking or parabolic interpolation.

Two more recent approaches include SWIPE (Camacho and Harris 2008) which performs spectral template matching of the input sound against sawtooth waveforms, and pYIN (Mauch and Dixon 2014), a probablistic variant of YIN with a Hidden Markov Model (HMM) as a post-processing step to predict probable pitch sequences. The pitch space of pYIN is divided into 480 bins ranging over four octaves from A1 (55Hz) to A5 (880Hz) in steps of 10 cents. Finally, CREPE (Kim et al. 2018) is the current state-of-the-art neural network for pitch tracking, and operates on the time-domain waveform directly, outputting a probability for 360 possible pitches ranging over six octaves between C1 (32.70Hz) and B7 (1975.5Hz) in steps of 20 cents. CREPE describes previous approaches as DSP-pipeline and heuristic-based, while claiming to be the first data-driven pitch tracker. The results of CREPE show it beating pYIN, the previous state-of-the-art according to several survey papers (Babacan et al. 2013; Knesebeck and Zölzer 2010).

Another way to categorize pitch tracking algorithms is by the domain (time or frequency) in which they operate. Jouvet and Laprie (2017) provide yet another comprehensive survey of pitch tracking algorithms. Time domain methods include the previously mentioned autocorrelation function (ACF), YIN and RAPT, in addition to time domain excitation extraction based on a minimum perturbation operator (TEMPO; Kawahara et al. 1999; Kawahara, Estill, and Fujimura 2001). Frequency domain methods include the previously mentioned SWIPE. Finally, there are a set of algorithms combine both approaches such as Aurora (Sorin et al. 2004), NDF (Kawahara et al. 2005), and REAPER.[1]

1. https://github.com/google/REAPER

Bibliography

Babacan, Onur, Thomas Drugman, Nicolas d'Alessandro, Nathalie Henrich Bernardoni, and Thierry Dutoit. 2013. "A comparative study of pitch extraction algorithms on a large variety of singing sounds," 7815–7819. May. https://doi.org/10.1109/ICASSP.2013.6639185.

Boersma, Paul. 1993. "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound." In *Proceedings of Institute of Phonetic Sciences 17,* 97–110.

Camacho, Arturo, and John G. Harris. 2008. "A sawtooth waveform inspired pitch estimator for speech and music." *The Journal of the Acoustical Society of America* 124 (3): 1638–1652. https://doi.org/10.1121/1.2951592. eprint: https://doi.org/10.1121/1.2951592. https://doi.org/10.1121/1.2951592.

Cheveigné, Alain, and Hideki Kawahara. 2002. "YIN, A fundamental frequency estimator for speech and music." *The Journal of the Acoustical Society of America* 111 (May): 1917–30. https://doi.org/10.1121/1.1458024. http://audition.ens.fr/adc/pdf/2002_JASA_YIN.pdf.

Dubnowski, John J., Ronald W. Schafer, Senior Member, Lawrence R. Rabner, and Senior Member. 1976. "Real-time digital hardware pitch detector." *IEEE Trans. Acoust., Speech, Signal Processing,* 2–8.

Jouvet, Denis, and Yves Laprie. 2017. "Performance analysis of several pitch detection algorithms on simulated and real noisy speech data," 1614–1618. August. https://doi.org/10.23919/EUSIPCO.2017.8081482. https://hal.inria.fr/hal-01585554/document.

Kawahara, Hideki, Alain Cheveigné, Hideki Banno, Toru Takahashi, and Toshio Irino. 2005. "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," 537–540. January.

Kawahara, Hideki, Jo Estill, and Osamu Fujimura. 2001. "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis." *2nd MAVEBA, Firenze, Italy* (September).

Kawahara, Hideki, Haruhiro Katayose, Alain Cheveigné, and Roy Patterson. 1999. "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," vol. 6. January.

Kim, Jong Wook, Justin Salamon, Peter Li, and Juan Pablo Bello. 2018. *CREPE: A Convolutional Representation for Pitch Estimation.* arXiv: 1802.06182 [eess.AS]. https://arxiv.org/pdf/1802.06182.pdf.

Knesebeck, Adrian von dem, and Udo Zölzer. 2010. "Comparison of pitch trackers for real-time guitar effects," 266–269. September.

Mauch, M., and S. Dixon. 2014. "PYIN: A fundamental frequency estimator using probabilistic threshold distributions." In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 659–663. https://doi.org/10.1109/ICASSP.2014.6853678. https://www.eecs.qmul.ac.uk/~simond/pub/2014/MauchDixon-PYIN-ICASSP2014.pdf.

McDermott, Josh, and Marc Hauser. 2005. "The origins of music: Innateness, uniqueness, and evolution." *Music Perception - MUSIC PERCEPT* 23 (September): 29–59. https://doi.org/10.1525/mp.2005.23.1.29. https://web.mit.edu/jhm/www/Pubs/McDermott_2005_music_evolution.pdf.

McLeod, Philip, and Geoff Wyvill. 2005. "A smarter way to find pitch." January. http://www.music.mcgill.ca/~ich/research/misc/papers/cr1172.pdf.

Moore, Brian C. J. 2013. *An Introduction to the Psychology of Hearing* [in English]. 6th ed. 203–242. United Kingdom: Emerald Group Publishing Limited.

Noll, A. Michael. 1967. "Cepstrum Pitch Determination." *The Journal of the Acoustical Society of America* 41 (2): 293–309. https://doi.org/10.1121/1.1910339. eprint: https://doi.org/10.1121/1.1910339. https://doi.org/10.1121/1.1910339.

Plack, C.J. 2013. *The Sense of Hearing* [in English]. 2nd ed. 117–118. United Kingdom: Psychology Press Ltd.

Ross, M., H. Shaffer, A. Cohen, R. Freudberg, and H. Manley. 1974. "Average magnitude difference function pitch extractor." *IEEE Transactions on Acoustics, Speech, and Signal Processing* 22 (5): 353–362. https://doi.org/10.1109/TASSP.1974.1162598.

Sorin, Alexander, T. Ramabadran, Dan Chazan, Ron Hoory, M. Mclaughlin, D. Pearce, F.C. Wang, and Yaxin Zhang. 2004. "The ETSI extended distributed speech recognition (DSR) standards: client side processing and tonal language recognition evaluation," 1:I–129. June. ISBN: 0-7803-8484-9. https://doi.org/10.1109/ICASSP.2004.1325939.

Talkin, D. 2005. "A Robust Algorithm for Pitch Tracking (RAPT)."