An Iterative Approach to Pitch–marking of Speech Signals without Electroglottographic Data

Ulrich Germann

University of Toronto

Abstract

We propose an iterative approach to high–quality pitch–marking of speech recordings without the use of laryngographic data. Our method first identifies islands of pitch marks that can be determined with high confidence. These islands are then extended into neighboring regions. A second round of island identification and extension with lower quality requirements fills the remaining gaps. We evaluate this pitch–marking method against pitch–marks produced with the *Praat* sound analysis software [1].

1. Introduction

Pitch and duration modification of recorded speech signals is an important subtask in concatenative speech synthesis. *Overlap–and–Add* (OLA) techniques transform overlapping, tapered windows of the signal and then add the results of these transformations to obtain a pitch– and/or duration–modified signal. The result tends to be audibly better if these operations are synchronized with the pitch period. Since concatenative speech synthesis "glues together" different speech samples to generate new speech, we also need to align these samples with respect to their phase to avoid clearly perceptible phase shifts in the generated signal. The point of glottal closure is the conventional reference point for both pitch–marking, which is therefore also known as *Glottal closure instance* (GCI) detection.

Traditionally, the best results for automatic pitch–marking have been obtained with the help of an electroglottograph (EGG), a device that monitors the electrical conductivity of the neck during speech and is thus able to reveal the openings and closings of the glottis during voiced speech precisely. However, EGG data is not always available. Another reliable method of pitch marking is the manual annotation of the wave form — the human eye is quite good at spotting the periodically recurring patterns of troughs and peaks in the signal. As tedious and repetitive the task may be for the human annotator, automatic detection of GCIs has proven to be a difficult and challenging task. Numerous techniques and methods have been proposed [2, 3, 4, 5, 6, 7, 8, 9, 10, 11], but most of them turn out to work very well on some types of signals only and perform poorly on others. This paper presents an iterative approach to pitch marking designed to work reliably across a broad spectrum of speech signals. In essence, the method presented here first identifies regions that can be pitch–marked with high confidence and projects pitch information from this high–confidence area into adjacent voiced regions. Remaining gaps are then filled with the a second iteration of the same steps but lower confidence thresholds.

In the following, we first provide a brief overview of existing approaches to pitch–marking and GCI detection. Section 3 presents our approach, which is evaluated in Sec. 5.

2. Related work

There are two major approaches to pitch estimation in speech signals.

Correlation–based measures [2, 3, 12] look at wave form similarity. They try to find the offset $\delta > 0$ that maximizes the correlation between sample points at times t and $t + \delta$ over a certain interval. Conventionally [2], the autocorrelation of the signal s at time t at the lag (delay) m is defined as follows

$$\phi_t(m) = \sum_{n=0}^{N-m-1} \frac{(s[t+n]w[t+n])(s[t+n+m]w[t+n+m])}{N} \tag{1}$$

where w is a windowing function that gives different weigth to different samples in the analysis window. Maxima in the autocorrelation function indicate periodicity in the signal. The problem with this measure is that it underestimates the auto–correlation of the signal at longer delays m. Note that the denominator is given by the analysis window size N, not by the degrees of freedom. Thus, the function gradually tapers off towards zero, and peaks due to resonance may dominate over peaks at the true fundamental frequency, promting this measure to over-estimate the actual pitch occasionally (*pitch duplication*).

The *normalized cross-correlation* (NCC) measure

$$\alpha_t(m) = \cos(\theta) = \frac{\sum_{n=-N/2}^{N/2-1} s[t+n]s[t+n+m]}{\sqrt{\sum_{n=-N/2}^{N/2-1}(s[t+n])^2}\sqrt{\sum_{n=-N/2}^{N/2-1}(s[t+n+m])^2}} \tag{2}$$

avoids this problem but is prone to *pitch halving*. Due to averaging effects, the cross-correlation at a lag of two or three pitch periods can be higher than at a lag of one pitch period. Boersma [12] has shown that a normalizing the auto–correlation function with a Gaussian windowing function w by dividing it by the auto-correlation of the windowing function itself can lead to very accurate pitch estimates.

Other techniques [4, 5, 6, 7, 8, 9] try to determine prominent, periodically recurring events in the speech signal, usually the instance of glottal closure.

Strube [4], Wong et al. [5], and Ma et al. [6] investigate noise levels within the signal. It is based on the conjecture that during glottal closure, the glottal impulse can resonate in the vocal tract without perturbance by air streaming in from the lungs. Thus, the signal should show less short-term variance than during glottal opening. Methods in this family of pitch marking methods slide a short (2ms) analysis window across the signal and compute measures related to the autocovariance matrix of this analysis window. Extrema in the computed values (e.g., the determinant of the autocovariance matrix [4], or the Frobenius norm of a related matrix [6]) indicate the instance of glottal closure. These methods work well only for certain types of speech signals and are not robust against noise.

Kadambe & Budreaux-Bartels [7, 8] use wavelet transforms to detect instances of significant perturbance in the signal. They observe that the perturbance caused by glottal closure leads to high coefficents across multiple scales of a dyadic wavelet transform of the original speech signal. In a similar fashion, Tuan & d'Alessandro use cosine–based wavelets to perform band–pass filtering of the in multiple frequency ranges. "Lines of maximum amplitude" across the various filter outputs indicate instances of glottal closure. In an evaluation by Wendt & Petropolu, Kadambe & Budreaux-Bartel's method "produced almost perfect results for synthesized speech signals" but "when tested with real speech signals failed" [11].

Both correlation–based and event–spotting approaches can be combined with dynamic programming to find maximally periodic sequences of recurring salient events in the signal, e.g. high peaks in the wave form [13, 14].

3. Iterative pitch–marking

We now describe our iterative approach to GCI detection.

3.1. Finding pitch period candidates

We first perform dyadic low-pass filtering of the signal with cubic B-spline wavelets [15] to obtain a signal in the frequency range 0–500Hz and align lows in this signal with lows in the originaly signal by following lines of minumum amplitude, in analogy to the method presented in [9]. These lows constitute potential segment (pitch period) boundaries.

Let $\alpha_{t,d}$ be the normalized cross–correlation of the signal with an analysis window of length $2d$, centered at point t, at a phase shift (offset) of d. We select all segments (s_t, s_{t+d}) as initial pitch period candidates for which $\max(\alpha_{t,d}, \alpha_{t+d,d}) > .9$. Segment hypotheses that do not correspond to a fundamental frequency in the range 50–600Hz are not considered.

We then perform dynamic programming to find chains of segments with good and smooth pitch estimates. Each segment is represented in the search graph by a vertex with vertex cost

$$v_{t,d} = \frac{\alpha_{t,d} + \alpha_{t+d,d}}{2d} \tag{3}$$

The division by the segment length is meant to penalize long pitch estimates. We found that especially near transitions between phones, averaging effects often lead to significantly higher cross-correlation scores for segments of two or three pitch periods. Transition costs between vertices are given by the formula

$$t_{t,d,t',d'} = (\max(\frac{d}{d'}, \frac{d'}{d}) - 1)^2 \tag{4}$$

Transitions between segment hypotheses are allowed only if two conditions are met:

1. $t' = t + d$, that is, the two segments are immediately adjacent.
2. $\max\left(\frac{d}{d'}, \frac{d'}{d}\right) < 1.5$

The latter condition is meant to prevent jumps into higher or lower octaves.

3.2. Establishing GCI seed chains

The dynamic programming process leads to a set of partially overlapping chains of segment hypotheses. Within each chain, we identify sequences of peaks of high amplitude, exactly one in each segment, that are as evenly spaced as possible. Again, we do this by dynamic programming. Each peak (GCI candidate) corresponds to a vertex. Vertex costs are given by the amplitude score

$$n_i = \left(1 - \frac{A_i}{\max(A_{i-.7\pi_i} \ldots A_{i+.7\pi_i})}\right)^2 \tag{5}$$

where A_i is the interpolation of the respective peak's absolute amplitude and its amplitude relative to the amplitude of the closest preceding trough in a smoothed[1] version of the signal, and π_i the length of the segment hypothesis containing the peak. Transition costs are given by the squared difference between the expected distance between the two GCI candidates (based on the average length of the segment

[1] Smoothing is performed by low–pass filtering of the original signal.

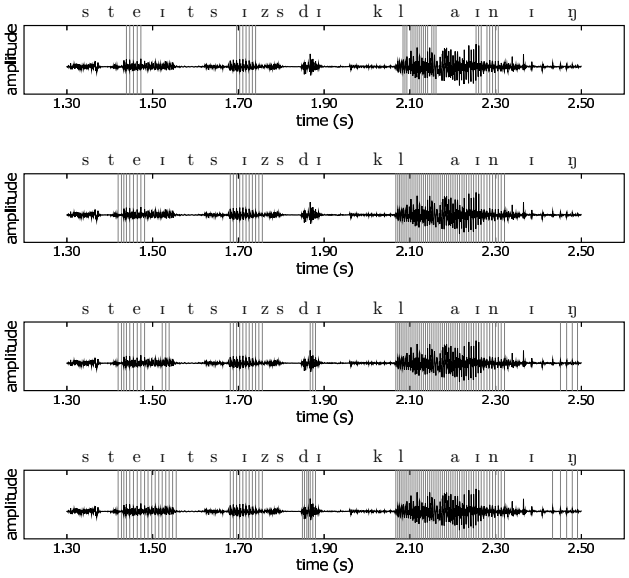

Figure 1: Pitch–marks for the speech sample "... States is declining." after each of the four processing stages. The four stages are described in the text.

candidates containing the two peaks) and the actual distance between the two peaks.

We then tile the resulting chains of GCI candidates over the signal in a best–first fashion. The chains are ranked by length (longest first); ties between chains of the same length are broken by comparing the respective path costs. We rank by path length rather than path cost to cope with short chains of accidental pitch–marks at peaks corresponding to the first and/or second formant, which occasionally leads to short chains with very low path costs. Lower-ranking chains of pitch–mark hypotheses are considered compatible with higher-ranking ones if they either do not overlap with them at all, or if they agree completely in the overlapping regions. Lower–ranking hypotheses incompatible with higher–ranking ones are discarded.

4. Pitch information projection

This first round of GCI estimation leads to islands of high-precision pitch–marks but low recall (see the top graph in Fig. 1). In the final step of each pitch–marking iteration, we extend these islands by projecting period information into adjacent regions for which no pitch–marks have been established yet. We detect gaps by comparing the distances between adjacent pitch–marks. Gaps are all regions where the distance between two adjacent pitch–marks is more than 1.6 times the expected distance.

At either edge of the gap, we select the peak in the region that is closest to where we would expect it, based on the period information available. Only peaks that have at least 80% of the maximum amplitude within a range of plus or minus half the expected period length are considered. Let d be the distance between the old frontier pitch–mark and the new pitch–mark candidate, and t the position of the old frontier pitch–mark. If the normalized cross–correlation between the signal segments in the ranges $[t-d:t-1]$ and $[t:t+d-1]$ is above a certain threshold (0.2 in the first iteration, 0.1 in the second), we accept the new pitch–mark. Otherwise, we reject it.

Figure 1 illustrates the entire pitch–marking process. First (top graph), areas with high normalized cross-correlation between segments are identified and pitch–marked with dynamic programming. Pitch information from these regions of high–confidence estimates is then projected greedily into adjacent regions (second graph from top). Stages 3 and 4 repeat these steps with lower cross-correlation thresholds.

Table 1: Precision and recall of the Praat pitch–marking algorithm and the algorithm presented in this paper (iter).

	male spkr.		female spkr.	
system	**prec**	**rec.**	**prec**	**rec.**
Praat	72.8%	79.4%	63.0%	68.3%
iter	71.3%	80.0%	70.6%	78.1%

5. Evaluation

We evaluated the performance of the pitch–marking method presented in this paper against pitchmarks produced by the sound analysis program *Praat* [1] on 10 short recordings each by a male and a female speaker from Bagshaw's database for the evaluation of pitch determination algorithms [16]. The data base consists of 50 short speech samples for a male and a female speaker as well as corresponding EGG data. We found that even with EGG data, producing a gold standard for the evaluation of pitch marking is not trivial. The most salient points in the EGG curve indicate glottal opening rather than glottal closure, so that there is always a (dynamically changing) delay between the most salient event in the EGG curve and the CGI mark produced by the pitch–marking algorithm. We used *Praat*'s pitch–marking function to extract glottal opening instances from the EGG data and manually verified them by visual comparison with both the EGG curve and the speech wave form. We then aligned the pitch marks proposed by each pitch–marker with the time stamps of the glottal opening instances extracted from the EGG data. A pitch mark p_i and a glottal opening time stamp t_j are considered aligned if $i = \mathrm{argmin}_{i'}\,\mathrm{abs}(p_{i'} - t_j) \wedge j = \mathrm{argmin}_{j'}\,\mathrm{abs}(p_i - t_{j'})$ $\wedge t_j < p_i$. Furthermore, we considered the distance between the pitchmarks produced by each system and considered the

one closer to the glottal opening instance the correct one. In the vast majority of the cases, both pitch markers agreed on pitchmarks within a window of 0.2 msec. Otherwise, the one farther from the glottal opeing time stamp was considered incorrect. Table 1 summarizes the results. An preliminary qualitative analysis of the pitch–marks produced by each system suggests that our pitch–marker is unfortunately still prone to pitch–halving, whereas the Praat pitch marker occasionally does not produce any pitch marks at all, or tends to put pitch marks on formant peaks rather than the peaks corresponding to GCIs.

6. Conclusions

We have presented a procedure for pitch–marking of human speech that combines several techniques to extract accurate pitch–mark information from natural speech recordings while maintaining robustness against variations in pitch and gender of the speaker. It appears to work well for low– and high–pitched voice as well as for male an female speech.

Our experiments in pitch–marking suggest that there is no one–size–fits–all pitch–marking method that performs accurately across wide ranges of signals. Adaptive algorithms that tune themselves to the speech signal at hand locally while keeping track of the big picture globally (through dynamic programming) show the greatest promise. Our experiments have also shown that it is feasible to locate glottal closure instances in speech signals without the help of laryngographic data.

7. References

[1] P. Boersma, "Praat: Doing phonetics by computer (version 4.4.16)." `http://www.praat.org/`, 2001. as of April 1, 2006.

[2] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, pp. 24–33, Febuary 1977.

[3] B. S. Atal, *Automatic Speaker Recognition Based on Pitch Contours*. PhD thesis, Polytechnic Institute of Brooklyn, 1968.

[4] H. W. Strube, "Dertermination of the instance of glottal closure from the speech wave," *Journal of the Acoustical Society of America*, vol. 55, no. 5, pp. 1625–1629, 1974.

[5] D. Y. Wong, J. D. Markel, and A. H. Gray, Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 350–355, Aug. 1979.

[6] C. Ma, Y. Kamp, and L. F. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal," *IEEE Transactions on Speech Audio Processing*, vol. 2, no. 2, pp. 258–265, 1994.

[7] S. Kadambe and G. F. Boudreaux-Bartels, "A comparison of wavelet functions for pitch detection of speech signals," in *International Conference on Acoustics, Speech, and Signal Processing*, (Toronto, Canada), pp. 449–452, May 1991.

[8] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detecton of speech signals," *IEEE Transactions on Information Theory*, vol. 38, pp. 917–924, Mar. 1992.

[9] V. N. Tuan and C. d' Alessandro, "Robust glottal detection using the wavelet transform," in *Eurospeech '99*, (Budapest, Hungary), 1999.

[10] J. D. Wise, J. R. Caprio, and T. W. Parks, "Maximum likelihood pitch estimation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, pp. 418–423, Oct. 1976.

[11] C. Wendt and A. P. Petropulu, "Pitch determination and speech segmentation using the discrete wavelet transform," in *IEEE International Symposium on Circuits and Systems*, (Atlanta, GA), pp. 45–48, May 1996.

[12] P. Boersma, "Accurate short–term analysis of the fundamental frequency and the harmonics–to–noise ratio of a samples sound," in *Proceedings of the Institute of Phonetic Sciences*, vol. 17, (University of Amsterdam, Netherlands), pp. 97–110, 1993.

[13] A. Kounoudes, P. A. Naylor, and M. Brookes, "The dypsa algorithm for estimation of glottal closure instants in voiced speech," in *International Conference on Acoustics, Speech, and Signal Processing*, (Orlando, Florida), pp. 349–352, May 2002.

[14] J.-H. Chen and Y.-A. Kao, "Pitch marking based on an adaptable filter and a peak-valley estimation method," *Computational Linguistics and Chinese Language Processing*, vol. 6, no. 5, pp. 1–12, 2001.

[15] W. Sweldens and P. Schröder, "Building your own wavelets at home," in *Wavelets in Computer Graphics*, pp. 15–87, ACM SIGGRAPH Course notes, 1996.

[16] P. C. Bagshaw, S. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," in *European Conference on Speech Communication and Technology (Eurospeech '93)*, (Berlin, Germany), pp. 1003–1006, Sept. 1993.