Accurate Pitch Marking for Prosodic Modification of Speech Segments

Thomas Ewender and Beat Pfister

Speech Processing Group
Computer Engineering and Networks Laboratory
ETH Zurich, Switzerland
{ewender,pfister}@tik.ee.ethz.ch

Abstract

This paper describes a new approach to pitch marking. Unlike other approaches that use the same combination of features for the whole signal, we take into account the signal properties and apply different features according to some heuristic. Basically we use a special type of energy contour for pitch marking. Where the energy information turns out to be not suitable as an indicator we resort to the fundamental wave computed from a contiguous F_0 contour in combination with detailed voicing information. Our experiments demonstrate that the proposed pitch marking algorithm clearly improves the quality of synthesised speech generated by a concatenative text-to-speech system that uses TD-PSOLA for prosodic modifications.

Index Terms: speech synthesis, glottal closure instants, pitch marks, automatic pitch marking, signal analysis.

1. Introduction

TD-PSOLA [1] is known to allow for high-quality pitch and time scale modification of speech segments for concatenative speech synthesis. The quality of TD-PSOLA-modified speech largely depends on how the speech signal is split into windowed double period segments. It is generally accepted and can easily be verified experimentally that the best results are achieved if double period segments start at a glottal closure instant (GCI) and end at the next but one GCI.

The GCI is commonly referred to as the maximum of the derivative of the airflow through the glottis. In terms of timing, the GCI is located towards the end of the closure of the glottis. It would be very easy to detect the GCI from the glottal flow signal. However, this signal is generally not available. Therefore the GCIs have to be estimated from the speech signal. These estimates are denoted here as pitch marks.

There are several known approaches to set pitch marks. One standard approach consists of two steps: First, pitch mark candidates are generated from local maxima either of the speech signal [2, 3, 4] or some wavelet components [5]. In the second step, a subset of these candidates is selected according to optimisation criteria that mainly account for the smoothness of the fundamental frequency contour [2, 3, 5, 6] and the waveform consistency of the signal around the pitch mark candidates [2, 6].

Another approach to pitch marking uses a pseudo-state space representation of speech frames and sets the pitch marks at the crossings of the trajectories with the Poincaré plane [7]. At the start of a voiced speech segment, this plane has to be placed according to some criterion, e.g. a local maximum of the signal or the point where the trajectories are most parallel. Both possibilities may result in pitch marks that are very bad estimates of the GCIs. A further problem of this approach is phase drift, i.e. in some signals the distances between the pitch marks may systematically be slightly too large or too small.

These methods are not robust enough for voiced segments with significant noise components as in voiced-to-unvoiced transitions or vice versa and in voiced fricatives. Furthermore, they are not designed to cope with irregular-pitched segments of vocal fry. Our new approach to pitch marking uses the maxima of the short-term energy contour as the main pitch marking criterion because the maximum of the derivative of the airflow through the glottis corresponds to the energy maximum in the speech signal. Our approach copes very well even with the difficult cases mentioned above. It will be outlined in the next section. In Section 3, the short-term energy and the fundamental wave will be introduced as the two basic features of our approach. Section 4 will illustrate some challenging cases, a detailed presentation of our approach follows in Section 5. Section 6 will describe our experiments and the results. In Section 7 some concluding remarks will be given.

2. Outline of new pitch marking

Estimating pitch marks from amplitude peaks often results in pitch periods with significant jitter that needs to be reduced by smoothing. We argue that since the concept of TD-PSOLA is based on the idea that the energy should concentrate in the centre of the windowed double period segments, pitch marks should be placed at peaks of the energy contour rather than at peaks of the signal. However, the short-term energy is not always a reliable criterion to set pitch marks (see Section 4). But it is possible to detect the locations where the short-term energy criterion should not be used. In these cases, the fundamental wave of the speech signal is used as a robust fallback feature to determine the pitch mark positions. It has to be emphasised that neither of the two features can be used alone for pitch marking. Some problematic cases for the short-term energy feature are illustrated in Section 4. Also the fundamental wave alone is not usable for setting pitch marks, because there is no fixed relation between the GCIs as estimated from the short-term energy and the phase of the fundamental wave (see Fig. 1). Hence a solution that is based solely on the fundamental wave would fail in some cases. Only with a combination of both features, as described below, good results may be attained for arbitrary voices.

3. The two main features for pitch marking

Our pitch marking algorithm is based on two features: the short-term energy contour and the fundamental wave. These features are extracted for each sampling point of the speech signal by means of an analysis window with the size of the local signal

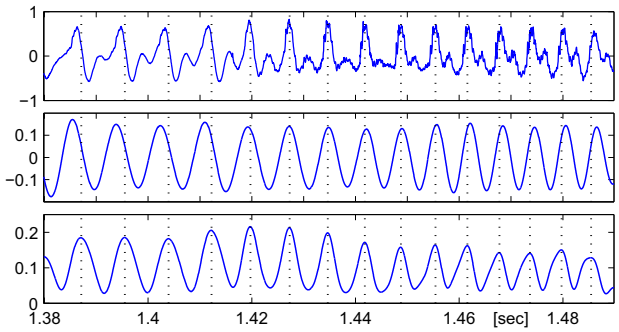

Figure 1: Speech signal (top), fundamental wave and short-term energy (bottom) of the phones [mɪ]. GCIs as estimated from the short-term energy are denoted as dotted lines. The phase of the fundamental wave at the estimated GCIs varies from 248° to 107°.

period. Thus the resulting features are continuous and smooth. The signal period is derived from a contiguous F_0 contour as resulting from the method described in [8].

3.1. Continuous short-term energy contour

The short-term energy for a sample i of a signal $x(\cdot)$ is computed as follows:

$$E(i) = \frac{\sum_j \left[x(j) \cdot w(j-i)\right]^2}{\sum_j w(j)^2}, \qquad (1)$$

where $w(\cdot)$ is a Hamming window of length T_0 centred at 0. The window of size T_0 is motivated by the fact that it is long enough to provide a good estimate for signals with noise components and short enough to provide the energy distribution within the length of one period. Since T_0 varies with time, the size of the Hamming window has to be adapted continuously, i.e., not only for each frame, but for each sample. Because a T_0 contour as it results from the optimisation described in [8] is specified with one value per frame, it has to be interpolated to get a T_0 value for each sample of the speech signal.

The resulting short-term energy contour is completely smooth and particularly shows no discontinuities at frame boundaries. A segment of such a short-term energy contour is shown at the bottom of Fig. 1.

3.2. Continuous fundamental wave

The fundamental wave can be computed by convolving the speech signal with a Hamming window of length T_0. This is a low-pass filter with a cut-off frequency of F_0 and zeros at the harmonics. The length of the Hamming window again has to be adapted continuously as described above. The resulting fundamental wave is also completely smooth and does not show discontinuities at frame boundaries. A segment of such a fundamental wave is shown in the middle plot of Fig. 1.

4. Problems with short-term energy

In most cases the short-term energy works fine for the pitch marks (see voiced frames of Fig. 4). However, there are cases where short-term energy peaks turn out to be not suitable for pitch marking. We illustrate such problematic cases below.

4.1. Pitch doubling

For nearly sinusoidal speech signals, basing pitch marks on short-term energy peaks would be problematic because the energy contour shows two almost equally high peaks per period (see Fig. 3 from 0.45 seconds). If the energy peaks are chosen without any further consideration this leads here to the well known effect of pitch doubling. This effect tends to appear more often with female voices where the fundamental frequency can be in the area of the first formant, leading to a dominating fundamental wave in the signal.

4.2. Jitter

The reason for jitter is that the energy contour has either not clear peaks or there are again two relative maxima per period. Fig. 2 shows such a speech signal. After 0.7 seconds the short-term energy contour first shows dual peaks and later broad peaks with the maximum changing from right to left and to right again. If these short-term energy peaks are used, the resulting pitch marks alternate between the left and right peaks which results in jitter.

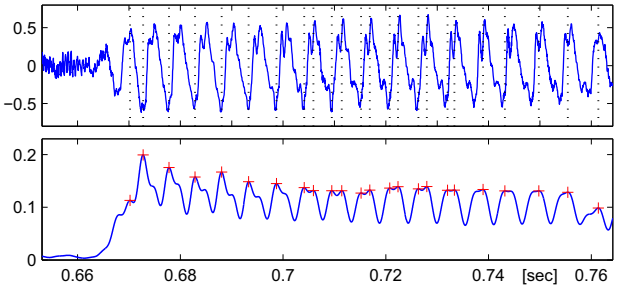

Figure 2: Speech signal (top) and short-term energy (bottom) of the phones [tã]. The dotted lines show the pitch marks if based solely on short-term energy peaks (marked with the plus signs), jitter occurs after 0.71 seconds.

4.3. Spurious energy peaks

High frequency noise can cause spurious energy peaks or can shift the energy peaks in the otherwise periodic signal. As can be seen in Fig. 3 at 0.30–0.36 seconds, the short-term energy contour exhibits irregular peaks, whereas the fundamental wave remains periodic and is not perturbed by the high frequency components. This is often observed with voiced fricatives next to vowels as it is the case in Fig. 3.

5. Combining short-term energy and fundamental wave

5.1. Reliability of the short-term energy

As is apparent from the previous section, a careful estimation of the short-term energy peaks' reliability as pitch mark indicators is most important. We used a set of 9 voices that were known to be problematic for pitch marking to establish the following criteria for determining unreliable short-term energy peaks:

- Prominence: the peak is not prominent enough compared to adjacent valleys (Fig. 3, from 0.45 sec.).
- Amplitude: the peak's amplitude is below a noise threshold (Fig. 3, until 0.345 seconds).

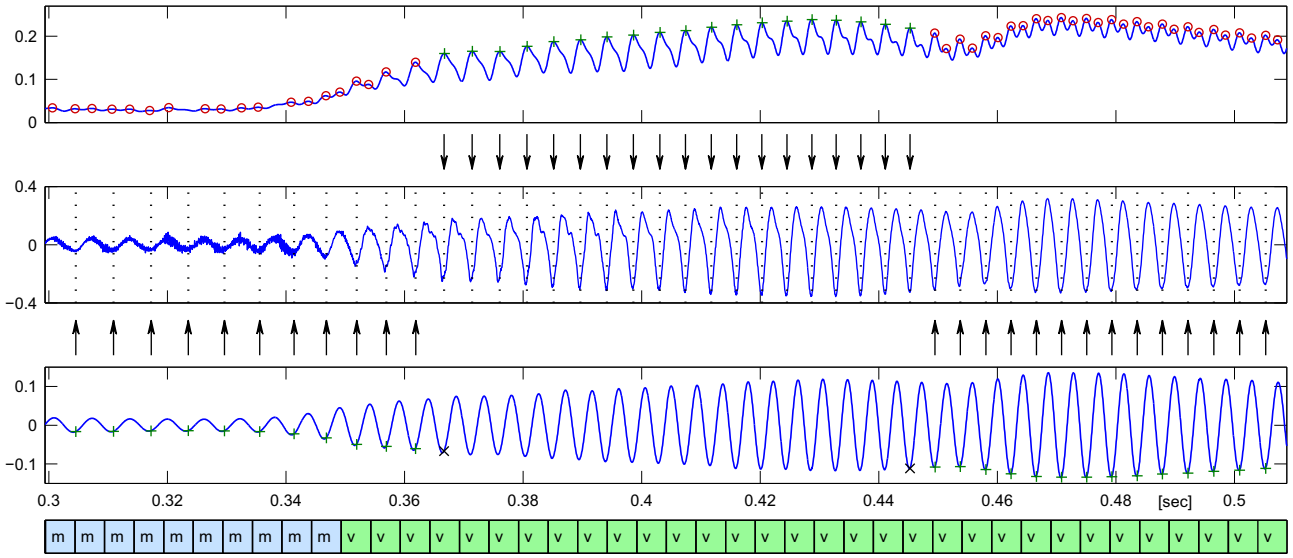

Figure 3: Speech signal of the phones [zum] from the word *zoomde* uttered by a female Dutch speaker (middle plot). On top the short-term energy is shown (for illustrative purposes we applied the square root), on the bottom the fundamental wave. The plus signs and circles designate extracted peaks. The peaks specified by plus signs are used as pitch marks, either directly as in the case of the short-term energy or in combination with the phase as in the case of the fundamental wave. The pitch marks are shown as dotted lines in the waveform signal. In the bottom plot, the voicing information is shown frame-wise, where the letters *v* and *m* denote *voiced* and *mixed* frames, respectively.

- Position of adjacent valleys: adjacent valleys are too close or too far away (Fig. 4, from 3.09 to 3.16 secs).
- Position of adjacent peaks: adjacent peaks are too close or too far away (Fig. 4, from 3.09 to 3.16 sec.).
- Form: the peak is too broad at a certain fraction of its maximum height (Fig. 2, from 0.69 sec.).
- Quality of neighbours: one of the adjacent peaks is considered an insufficient indicator by meeting one of the above mentioned criteria.

If at least one of these negative criteria is fulfilled, the energy peak is considered an insufficient indicator and the algorithm resorts to the fallback methods.

5.2. Reliability of the fundamental wave

In many cases where the energy contour shows no usable peaks because of the presence of high-frequency noise components (see e.g. the voiced fricative at the beginning of the signal in Fig. 3), the fundamental wave is almost perfectly periodic and does not exhibit any perturbation. In these cases the fundamental wave can be used as a fallback. However, there are cases where also the fundamental wave fails as a reliable pitch mark indicator, which is frequently the case with creaky voice segments (see Fig. 4 at 3.10–3.17 seconds). Therefore again a set of criteria for the detection of bad fundamental wave segments have been established:

- Amplitude: the amplitude is below a noise threshold (Fig. 4, from 3.1 to 3.15 sec.).
- Valley positions: valleys are too close or too far away from their adjacent peak (Fig. 4, from 3.1 to 3.165 sec.).
- Peak positions: peaks are too close or too far away from their neighbours (Fig. 4, from 3.09 to 3.155 sec.).
- Regularity of adjacent valleys: the ratio of the distances between the peak and its two adjacent valleys is too different from the local increase of F_0.

The fundamental wave is locally considered as not reliable, if at least one of the above criteria is fulfilled. The above mentioned noise thresholds are dynamically estimated for each signal using an energy-based loudness measure. The rest of the parameters was manually determined using a set of 9 voices and is virtually voice-independent.

5.3. Procedure to set the pitch marks

As input to the pitch marking procedure the short-term energy and the fundamental wave (as described in Section 3) are extracted from the signal, as well as frame-wise defined voicing information which distinguishes voiced, unvoiced, mixed-excitation, irregularly-glottalized and silence segments (see [8]).

We first select glottalized (voiced, mixed or irregular) segments that are delimited by unvoiced or silent regions. Every glottalized segment is processed as follows:

1. We use the time points of the reliable short-term energy peaks in the glottalized segment as initial pitch marks.
2. If there is a region in the segment where short-term energy peaks are unreliable we check if the fundamental wave can be used for setting additional pitch marks. If this is the case, we detect at which phase of the fundamental wave the last already set pitch mark has been set (marked with x-signs in Fig. 3) and interpolate or extrapolate along the reliable part of the fundamental wave[1].

[1] In the rare case when no reliable short-term energy peak was found in step 1, we select one of the short-term energy peaks that fulfil most of the reliability criteria.

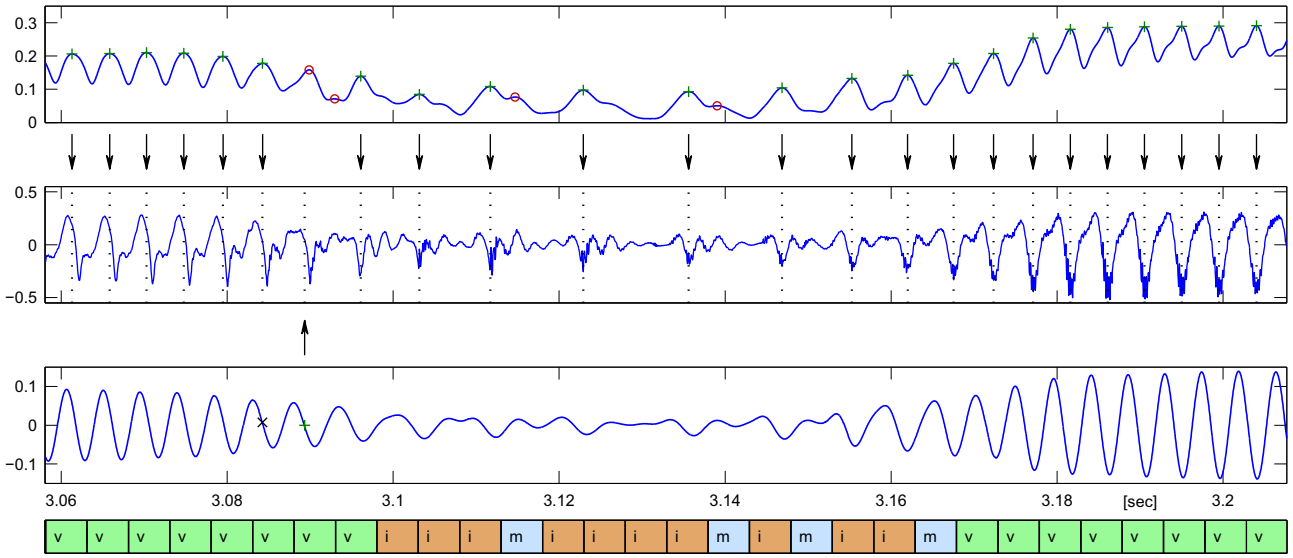

Figure 4: Speech signal of the phones [ŋiː] from the words *jiāng yī* uttered by a female Mandarin speaker with the same information as described in Fig. 3. The letters *v*, *i* and *m* denominate *voiced*, *irregular* and *mixed* frames, respectively.

3. For regions where still no pitch marks are set, we regard the voicing information: For regions classified as irregular we use prominent short-term energy peaks (which may be irregularly spaced) to set the pitch marks. For regions classified otherwise we set the pitch marks by interpolation (if there are pitch marks to the left and right of the region) or by extrapolation (at the start or the end of the glottalized segment) with a period length derived from the contiguous F_0 contour.

Finally, equidistant pitch marks are set in those segments that contain silent and unvoiced frames with some transition to the pitch mark distances at the borders of the glottalized segments.

6. Evaluation

Statistical evaluation can only be considered if evaluated pitch marks and reference pitch marks follow the same criteria. So we decided to evaluate our method through the performance of a diphone-based text-to-speech (TTS) system that applies TD-PSOLA. We built two systems each for four different voices, one male and one female Dutch voice, one female German and one male American English voice. The voices were recorded in studio quality with professional equipment. For the baseline system we used pitch marks generated with the cross-correlation algorithm from the Praat toolkit [9], for the other system we used our proposed method. With the exception of the difference in pitch marking the systems are otherwise equal.

The speech signals synthesised with our proposed method showed clear improvement for the two male voices, they contain less reverberation and sound less raspy. Most improvement was achieved for the bassy male American English voice, which sounds more sonorous and less frizzled. The female voices showed isolated improvements in mixed or irregular speech. We demonstrate the quality of our results by means of enclosed examples.

7. Conclusion

In this paper a new pitch marking method has been introduced. The algorithm places pitch marks basically at the peaks of the short-term energy contour, which is a good estimate of GCIs. For speech segments where the energy peaks are not suitable, we use fallback methods based on the fundamental wave or on the F_0 contour. The quality of our pitch marking is eliminating virtually all artifacts that are caused by other pitch markers such as Praat in the synthetic signal processed with TD-PSOLA.

8. Acknowledgements

This work was supported by the Swiss Innovation Promotion Agency CTI. We thank SVOX AG for providing multi speaker recordings which were helpful to test and improve our proposed method. We cordially thank Cédric Schaller who contributed to this work with his master's thesis.

9. References

[1] F. Charpentier and E. Moulines, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," in *Proceedings of Eurospeech*, Paris, September 1989, pp. 13–19.

[2] V. Colotte and Y. Laprie, "Higher precision pitch marking for TD-PSOLA," in *Proceedings of the European Signal Processing Conference*, Toulouse, 2002, pp. 419–422.

[3] C.-Y. Lin and J.-S. R. Jang, "A two-phase pitch marking method for TD-PSOLA synthesis," in *Proceedings of Interspeech/ICSLP*, Jeju Island (Korea), October 2004, pp. 1189–1192.

[4] B. Kotnik, H. Höge, and Z. Kacic, "Noise robust F0 determination and epoch-marking algorithms," *Signal Processing*, vol. 89, no. 12, pp. 2555 – 2569, 2009.

[5] M. Sakamoto and T. Saitoh, "An automatic pitch-marking method using wavelet transform," in *Proceedings of Interspeech/ICSLP*, Beijing, September 2000.

[6] R. Veldhuis, "Consistent pitch marking," in *Proceedings of Interspeech/ICSLP*, Beijing, September 2000, pp. 207–210.

[7] M. Hagmüller and G. Kubin, "Poincaré pitch marks," *Speech Communication*, vol. 48, no. 12, pp. 1650–1665, 2006.

[8] T. Ewender, S. Hoffmann, and B. Pfister, "Nearly perfect detection of continuous F0 contour and frame classification for TTS synthesis," in *Proceedings of Interspeech*, Brighton, September 2009, pp. 100–103.

[9] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2002.