An Efficient and Robust Pitch Marking Algorithm on the Speech Waveform for TD-PSOLA

Aimilios Chalamandaris [#1*1], Pirros Tsiakoulis [#2*2], Sotiris Karabetsos [#3*3], Spyros Raptis [#4*4]

[#] *Institute for Language and Speech Processing- Athena Research Centre*
Artemidos 6 & Epidavrou, 15125 Athens, Greece
[1] achalam@ilsp.gr, [2] ptsiak@ilsp.gr, [3] sotoskar@ilsp.gr, [4] spy@ilsp.gr

[*] *innoetics ltd., Knowledge and Multimodal Interaction Technologies*
Artemidos 6 & Epidavrou, 15125 Athens, Greece
[1] aimilios@innoetics.com, [2] ptsiak@innoetics.com, [3] sotoskar@innoetics.com,
[4] sraptis@innoetics.com

***Abstract*— In a Text-to-Speech system based on time-domain techniques that employ pitch-synchronous manipulation of the speech waveforms, one of the most important issues that affect the output quality is the way the analysis points of the speech signal are estimated and the actual points, i.e. the analysis pitchmarks. In this paper we present our methodology for calculating the pitchmarks of a speech waveform, a pitchmark detection algorithm, which after thorough experimentation and in comparison with other algorithms, proves to behave better with our TD-PSOLA-based Text-to-Speech synthesizer (Time-Domain Pitch-Synchronous Overlap Add Text to Speech System).**

I. Introduction

In recent years Text-to-Speech (TTS) systems have known a dramatic improvement offering a near-natural synthetic speech. This improvement was driven by the growth of the available storage and computational power, making the use of large speech repositories feasible for TTS via the unit selection algorithm [1]. Such TTS systems usually rely on the TD-PSOLA algorithm [2], which allows us to concatenate speech waveforms and manipulate their prosodic characteristics, without much computational effort and with affordable distortion levels.

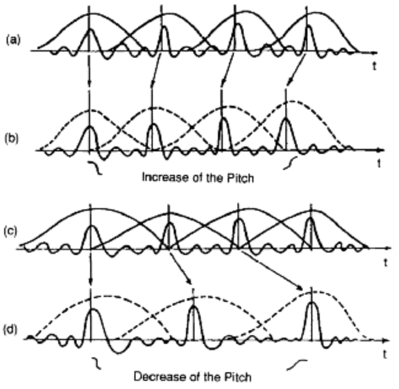

Fig. 1. A schematic illustration of the TD-PSOLA method by concatenation and overlap add of the ST-signals. (Source: [3])

A PSOLA-based TTS system uses short-term signals (ST-signals) from almost naturally uttered pre-recorded speech as the initial material that concatenates in a way that allows modifying pitch and duration of the signal. The ST-signals are usually two pitch-period long windowed signal frames, centered at specific points which are called analysis pitchmarks (Fig. 1). The distance between two consecutive pitchmarks reflects the local pitch level of the speech waveform, and their accurate and consistent calculation is a decisive factor of the synthesized speech quality. In fact, the more inconsistent the pitchmarks are the hoarser and eventually more robotic will the synthetic speech sound [4]. This is actually one of the main characteristics and limitations of the TD-PSOLA algorithm.

II. Pitch-marking algorithms

There have been several algorithms for estimating the pitchmarks for a TTS database, such as [5], [6] and [7], most of which utilize the EGG signal of a laryngograph in order to avoid spurious pitchmarks. Although those methods are accurate and able to determine glottal closure and opening phases of the speech signal (GCI and GOI instances), they suffer from a severe drawback: they require the synchronized recording of a laryngograph.

Other approaches and realizations of TTS systems admit that they manually identify the analysis pitchmarks in order to overcome errors introduced by automatic algorithms. However, manual annotation of the pitchmarks is a labor and time-consuming task and often prone to errors if the initial recordings of a TTS system are several hours long.

As for automatic approaches for estimating the analysis pitchmarks based on the speech signal, there have been also several approaches, others aiming to peaks or valleys in the waveform [8], and others aiming to identify the glottal-closure instances (GCI's) of the voice source signal [9]. Most of them work well with good accuracy and based on sophisticated Dynamic Programming techniques, and they manage to address the problem they try to answer adequately. However, in the case of a TTS system's database, experimentation has shown that what is of the highest importance, as far as the pitchmarks are concerned, is their consistency among different recordings of the same speaker.

III. The Proposed Algorithm

Our approach addresses the issue of pitch marking in the specific framework of a TTS system and consists of a complete mechanism for identifying accurately and efficiently the analysis pitchmarks for use in a TD-PSOLA TTS system. It is based solely on the speech signal and it does not require the EGG signal of the recording.

A. Overview

The approach we propose is applied on the speech waveform and is based on a dynamic programming methodology for identifying the analysis pitchmarks through a priori knowledge of the pitch contour. The flowchart of the algorithm is illustrated in figure 2, where one can actually identify the different stages. Detailed description of every stage is given in the next paragraphs, as well as an overview of the results that our method produces.

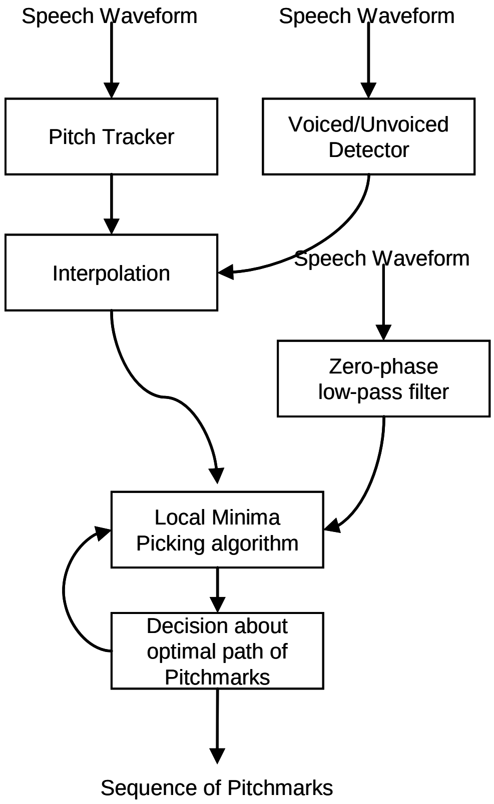

Fig. 2. The flowchart of our approach for pitchmarks' detection on the speech waveform.

1) Pitch Contour Tracking

The first stage of the algorithm involves the accurate pitch contour tracking of the speech waveform. Although this stage is often prone to introducing errors due to noise or disordered speech, which would consequently affect the immediately following process, in the case of recordings destined to consist the primary material for a database of a TTS system, these effects do not exist, since the recording is performed in a noise-proof studio and the speaker is a professional voice talent, with trained voice. Nevertheless, in order to achieve optimal pitch tracking with as few errors as possible, we use a 3-tier pitch tracker which manages to eliminate erroneous estimations by applying a majority decision. Three different pitch trackers are used, namely an autocorrelation pitch tracker as described in [10], a forward cross correlation one [11], and a pitch tracker based on zero-phased low pass filtering [12].

2) Voiced/Unvoiced Detection

In order to eliminate spurious pitch contour detection in unvoiced segments of the waveform, we employ a intermediate stage of voiced/unvoiced detection. As parameters for the voiced/unvoiced decision we use the local pitch variation, the intensity level and the autocorrelation coefficient. By doing so we identify the isles of confidence for voiced parts of the speech signal as well as the corresponding pitch contours.

3) Pitch Interpolation

As a second phase of our methodology we chose to interpolate the pitch contour in the unvoiced parts of the signal, in order to acquire a continuous pitch contour for the entire speech waveform. This is a necessary step in order to calculate pitchmarks in the unvoiced segments. Although there is no natural meaning in the pitchmarks located in the unvoiced parts of the signal, they actually serve an important purpose in the TTS system, since they compensate for false voiced/unvoiced detections of the pitch tracker and for errors occurred in the segmentation process of the speech signal into phones or diphones.

4) Pitchmarks Selection

The final stage of our pitch marking algorithm is the pitch-synchronous detection of the local minima in the low-pass filtered speech signal. The speech signal is low-pass filtered through an FIR zero-phase filter, in order to suppress high-frequency components that would be able to introduce errors in the minima-picking algorithm. The local-minima picking algorithm identifies the strongest local minimum in every voiced interval in the filtered signal and identifies iteratively the neighboring local minima, towards both ways, right and left, on a pitch synchronous basis, according to the pre-estimated pitch contour. The process is as follows:

- Let t_1 be the first absolute minimum of a voiced isle, in the space $\left[t_{start}+\frac{T_0}{2}, t_{end}-\frac{T'_0}{2}\right]$ where t_{start} and t_{end} are the starting and ending boundaries of the voiced isle respectively and T_0 and T'_0 are the periods at those boundaries, as interpolated by the pitch contour on those points.
- From this point on, we recursively search for point t_i towards t_{start} and t_{end} which correspond to the absolute minima in the space $\left[t_{i-1}-1.2\cdot T_0, t_{i-1}-0.8\cdot T_0\right]$,

where T_0 is the local interpolated period by the pitch contour.

The same process is performed for several different starting points (significant local minima) for every voiced interval, and the sequence of pitchmarks which provides the best concordance with the estimated pitch contour and the lowest local minima is picked.

In the unvoiced parts of the signal the pitchmarks are estimated explicitly without identifying the local-minima of the signal, but simply pitch synchronously, via the interpolated pitch contour.

In the following section we present the results of our methodology and of the experiments we carried out in order to further asses our algorithm.

IV. Results

The results of the proposed algorithm are significantly accurate and consistent, a feature that is necessary in this field of application. In order to assess our approach, aside from manual visual assessment, we carried out two different experiments: one experiment in comparison to the EGG signal of recorded speech, and another one where we carried out a MOS (subjective assessment) of the TTS output employing different pitch marking algorithms.

B. Visual Assessment

The results of our approach are more consistent than other approaches, such as the DYPSA algorithm, which although it performs very well in general and manages to identify the GCI and GOI points of the speech signal with good accuracy, it may produce phase mismatches in the location of the pitchmarks throughout the entire set of the recordings, and this would introduce hoarseness in the synthesized speech.

With the methodology we propose the analysis pitchmarks are consistently located at the same relative position of every pitch period of the speech signal. These pitchmarks, even though they are not exactly at the CGI points of the speech signal, they are consistently located at an offset of those instances even in rapid pitch variations (fig. 5). This consistence is one of the main features our TTS systems benefits from. This is a cross-speaker characteristic since the same algorithm provides the same results in different speakers.

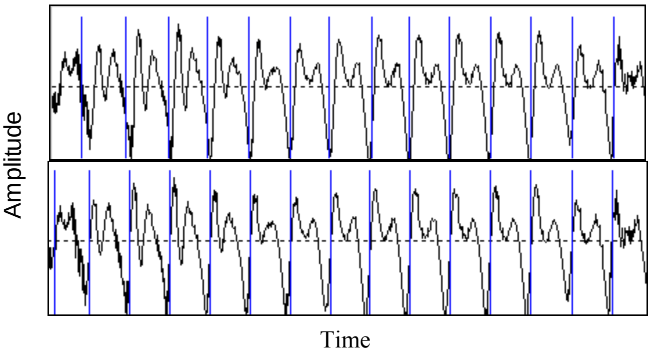

Fig. 3. (a) pitchmarks produced by the DYPSA algorithm, and (b) the pitchmarks produced by our algorithm. The second graph illustrates more consistent pitchmarks, with better application in TTS systems.

C. Assessment in Reference to EGG signal

In order to investigate the behaviour of this algorithm in reference to the actual GCI points of a speech waveform, as those were derived from the corresponding EGG signal, we carried out an experiment with the recordings of 6 different speakers, 3 males and 3 females, and we compared the produced pitchmarks with the actual Glottal Closure Instances of the waveform. As already mentioned, the pitchmarks produced by our algorithm were consistently placed at a very small offset of the vicinity of the actual GCI of the waveform.

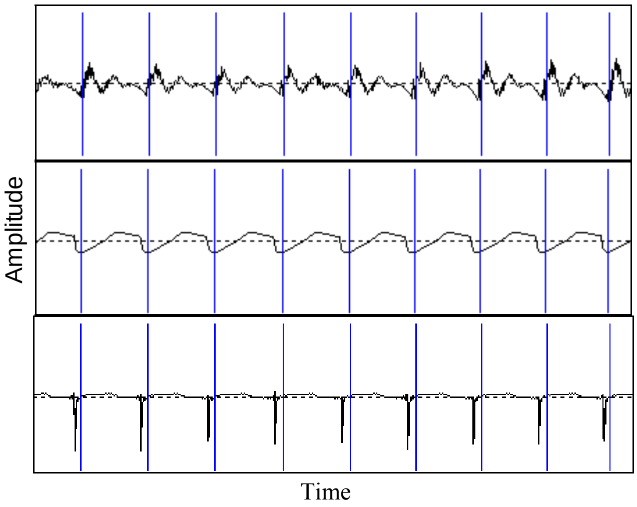

Fig. 4. (a) The pitchmarks produced by the proposed algorithm (b) in reference with the corresponding EGG signal, (c) in reference with the derivative of the EGG signal.

We also discovered that the average offset distance between the pitchmarks and the corresponding GCI points, depends on the speaker, something however that does not affect the quality of the TTS system, since the latter manipulates the signal of a single speaker each time and any inconsistency observed among different speakers does not affect the actual TTS system.

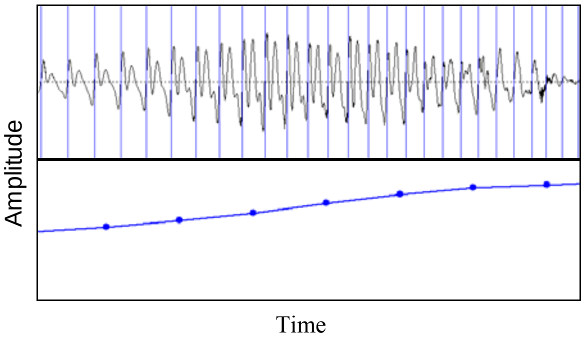

Fig. 5. Pitchmarks produced by the algorithm on a waveform where there is rapid pitch variability. (a) speech signal with pitchmarks, (b) pitch contour.

In figures 6 and 7, one can observe the different distributions of the measured detection accuracy of the proposed algorithm, with reference to the actual GCI points.

In different speakers the distribution is also different; nevertheless it can be identified as a mixture of two Gaussians, one with very small variance which covers almost all the data and the other with a larger variance but little coverage mainly in the voiced/unvoiced boundaries. This provides a clear idea of the consistency of the pitchmarks throughout the recordings of the same speaker.

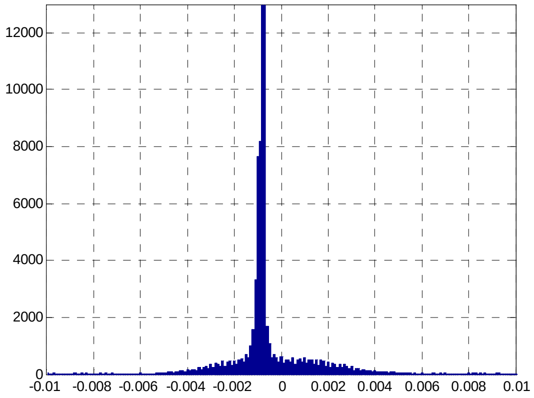

Fig. 6. The distribution of the measured detection accuracy compared to the original EGG signal. X-axis illustrates time in seconds. Graph for male speaker KED. Source CMU KED Database.

The original GCI points were extracted from the derivative of the EGG signal automatically and were manually assessed on a small random sample.

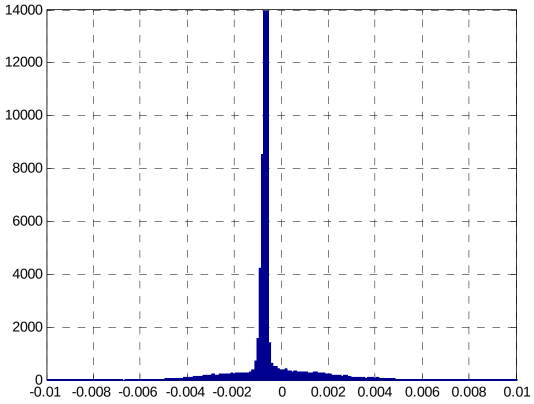

Fig. 6. The distribution of the measured detection accuracy compared to the original EGG signal. X-axis illustrates time in seconds. Graph for female speaker. Source TIMIT Database.

The above results will be further investigated in future work in order to identify potential ways of optimizing the accuracy of the proposed method.

D. MOS Experiment Assessment

In order to further investigate the appropriateness of our approach for the field of Text-to-Speech application, we carried out an objective assessment of the output of our TTS system incorporating different databases each time, as far as the analysis pitchmarks of the speech signal are concerned. Our TTS system is a TD-PSOLA-based unit-selection system, supporting the Greek language and it produces near-natural synthetic speech, with a database containing over than 8 hours of natural speech, appropriately segmented and annotated. More specifically, we developed three different databases of the same speaker where the only difference among them was the algorithm we used for the pitchmarks' location. For the pitchmarks' identification we used the proposed method, the DYPSA algorithm, which we have assessed and it works well with TTS systems, and the pitchmarks produced by the Praat programming environment [13] using the cross-correlation coefficients of the signal. In the unvoiced intervals of the speech signal, in the cases of DYPSA and Praat, we introduced "dummy" pitchmarks at a rate of the average level of the local pitch, as measured in the neighbouring voiced intervals, in order to avoid discontinuities when manipulating the signal in the voiced/unvoiced boundaries. We synthesized 20 different stimuli, each of them three times (one for every different algorithm) and we asked 13 different listeners, (all of them were speech experts with long experience in TTS and ASR technologies), to grade the speech quality of the stimuli, only as far as signal discontinuities and the level of signal's hoarseness were concerned. The stimuli were presented in a shuffled order and the subjects did not have knowledge of the underlying differences among the stimuli. The subjects were asked to grade the stimuli with a grade from 1 to 5, with 1 meaning very bad and 5 meaning best.

Our experiment showed that our approach performed better in comparison to the other approaches, when used in our TTS system.

TABLE 1: THE MOS EXPERIMENT RESULTS: THE PROPOSED METHOD VS DYPSA AND PRAAT-BASED PITCH MARKING.

	MOS (1-5 Grade)	St. Deviation
Our approach	4.22	0.44
DYPSA	4.11	0.60
PointProcess (cc) Praat	3.33	0.50

As one can observe in Table 1, the stimuli produced by the TTS which incorporated the analysis pitchmarks derived from the proposed algorithm received the highest grade, leading to the conclusion that it performs better in the application framework of TTS systems, offering a useful improvement compared to other methods.

Additionally to the MOS assessment experiment we carried out a subjective preference experiment in which the listeners were given 20 different triads of stimuli in a shuffled order, and they were asked to prioritise them according to their preference. In the following table one can observe the results for each pitch marking algorithm.

TABLE 2: THE RESULTS OF THE PREFERENCE EVALUATION OF THE STIMULI.

	Preference %
Our approach	53.8
DYPSA	38.5
PointProcess (cc) Praat	7.7

It is obvious that the stimuli derived with the pitchmarks from our method rank higher in the subjective assessment test, validating our hypothesis that our method offer improved results compared to the other methods.

V. Conclusion – Discussion

Our approach for identifying the analysis pitchmarks for a PSOLA-based TTS system is based on an algorithm that aims to place the pitchmarks in such way that they are consistent throughout the entire set of recordings of the same speaker. Although it does not achieve to identify the GCI points of a speech signal precisely, it manages consistently and robustly to place the pitchmarks at a constant offset distance in the vicinity of the GCI points. By doing so it helps avoid phase discontinuities during the process of the concatenation of the ST signals by overlap add. It helps the TTS system provide high quality synthetic speech with low signal distortion.

Algorthms such as DYPSA perform very well with success rates of 80-90%, however they seem to be slightly inferior for use with a TTS system since any inconsistence they produce will lead to deterioration of the quality of the synthesized speech.

VI. References

[1] A. Hunt, and A. Black, (1996). Unit selection in a concatenative speech synthesis system using a large speech database Proceedings of ICASSP 96, vol 1, pp 373-376, Atlanta, Georgia.

[2] F. Charpentier and E. Moulines, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” in Proc. EUROSPEECH, vol. 2, 1989, pp. 13–19.

[3] S. Lemmetty, Review of Speech Synthesis Technology, Master’s Thesis, Helsinki University of Technology, 1999

[4] D. O'Brien and A.I.C.Monaghan, "Concatenative Synthesis Based On A Harmonic Model" in IEEE Transactions On Acoustics, Speech, And Signal Processing, Vol 9, No. 1, January 2001, pp. 11 - 20.

[5] H.W. Strube, “Determination of the Instant of Glottal Closures from the Speech Wave,” J. Acoust. Soc. Am., vol. 56, pp. 1625-1629, 1974.

[6] A.W. Black, K.A. Lenzo,: Building Synthetic Voices, Language Technologies Institute, Carnegie Mellon University and Cepstral LLC. Retrieved from: http://festvox.org/bsv/ (2003/

[7] P. Taylor, A. Black, and R. Caley. The architecture of the festival speech synthesis system. In Proc. of the 3rd ESCA Workshop on Speech Synthesis, pages 305–310, 1998

[8] 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.

[9] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, “Estimation of glottal closure instants in voiced speech using the DYPSA algorithm,” IEEE Trans. Audio, Speech and Language Processing, vol. 15, no. 1, pp. 34–43, Jan. 2007.

[10] P. Boersma, (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam 17: 97-110.

[11] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition. Prentice Hall, 1993.

[12] I. Dologlou, G. Carayannis: "Pitch Detection based on zero-phase filtering", Speech Communication, Vol. 8, No 4, December 1989, pp. 309-318.

[13] P. Boersma, (1997). Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat/.

[14] M. Huckvale, Speech Filing System: Tools for Speech Research, University College London, 2000, [Online] http://www.phon.ucl.ac.uk/resource/sfs/.

[15] A. Black, P. Taylor, R. Caley, 1998. The Festival Speech Synthesis System. http://festvox.org/festival