Abstract: In this paper a Poincaré approach to pitch mark determination is presented. While speech has been interpreted in terms of nonlinear systems theory for quite some time, not much effort has been made to exploit this knowledge in the problem of pitch mark detection. This algorithm uses nonlinear state space embedding and calculates the Poincaré section at a chosen point in state space, pitch-marks are then found at the crossing of the trajectories with the Poincaré plane. The procedure is performed frame-wise to account for the changing dynamics of the speech production system. First results show promising performance, comparable to the pitch marking algorithm used in 'Praat', and outperforming it in case of irregular voices.

Keywords: Dysphonic speech, state-space-embedding, Poincaré section, pitch-marks.

I. INTRODUCTION

For pitch–synchronous processing of speech, accurate pitch-marks are essential. A particular challenge is the correct determination of pitch-marks for dysphonic voices. On the other hand, having a reliable method for pitch marking available, this could be used for enhancement of rough pitch, by reducing the fluctuations of the fundamental period. Accurate and robust methods for pitch detection are of interest for the analysis of dysphonic voices [1] and, e.g., for the measurement of jitter, methods to reliably determine the instantaneous fundamental period are necessary.

The nonlinear nature of the speech signal has been of increasing interest for several years now, starting in the early nineties [2].

Conventional algorithms, such as correlation based methods, assume linear models of speech production, though even for normal voices those models cannot fully explain the properties of the signal. For dysphonic speech, those models more or less fail due to the higher dimensional non-linearity inherent in the system. Especially, for strongly irregular voices, conventional algorithms for pitch mark determination fail and, therefore, the need for new methods is at hand. Non-linear methods seem to be a promising way of overcoming the weaknesses of the currently used approaches.

State-space approaches for dysphonic voice analysis have been proposed recently [3], [4]. Voice irregularities have been treated with nonlinear methods before, e.g., by performing noise reduction in state space [5].

The paper is organized as follows. Section II will give some background and review existing algorithms for pitch determination in state space. In section III the proposed state-space approach for pitch marking will be introduced and the algorithm will be explained. Section IV will show some results and finally section V will conclude the paper with a summary and an outlook.

II. BACKGROUND AND RELATED WORK

A non-linear dynamical system can be embedded in a reconstructed state-space by the methods of delays. According to Takens [6], the state space of a dynamical system can be topologically equivalently reconstructed from a single observed one-dimensional system variable. For a D-dimensional attractor it is sufficient to form a M ≥ 2D + 1 state space vector. The M-dimensional trajectory is formed from a speech signal vector x(n) by delayed versions of the signal x(n),

\[ x(n) = [x(n), x(n - \tau_d), ..., x(n - (N + 1)\tau_d)], \]

where \( \tau_d \) is the delay time, which has to be chosen to optimally unfold the attractor. If one chooses an arbitrary point on the attractor in an M-dimensional space then one can create a hyper-plane which is orthogonal to the flow of the trajectory at the chosen point. This is called the Poincaré plane. All trajectories, that return to a certain neighborhood of the initial point, cross the hyperplane and can be represented in dimension M − 1 compared to the original trajectory.

In 1997 Kubin [7] first suggested to use those Poincaré sections for the determination of pitch-marks and mentioned special applications for signals with irregular pitch period. Experiments showed very promising results for an example with vocal fry, where the pitch period doubles for some time. The pitch period was followed correctly.

Later Mann and McLaughlin [8] further worked with Poincaré maps and applied them to epoch marking for speech signals. They again saw promising results, but reported inability to resynchronize after, e.g., stochastic portions of speech.

More recently Terez [9] introduced another state space approach to pitch detection, using space-time separation histograms. Each point on the trajectory in state space is separated by a spatial distance \( r \) and a time distance \( \Delta t \). One can draw a scatter plot of \( \Delta t \) versus \( r \) or,
for every time distance, count the pairs within a certain neighborhood \( r \). This can then be normalized to yield a histogram (fig. 1). In case of periodicity in the signal, the histogram concentrates at certain \( \Delta t \) values, whereas others have rather low values. The first maximum of the histogram indicates the fundamental pitch period. Compared to the autocorrelation function the peak is much more significant and, therefore, offers improved performance. In case of noise-like signals the histogram is more evenly spread over all time distances. Since histograms are based on averaging statistics, localized pitchmarks cannot be determined reliably with this approach.

III. Description the Algorithm

Our algorithm builds on the before mentioned approaches. The algorithm works on a frame-by-frame basis to handle the changing system parameters.

For pitch mark detection the low-dimensional characteristics of the signal need to be observed. So the noise has to be removed, otherwise the attractor is hardly visible with 3-dimensional embedding (fig. 2). If the embedding dimension is high enough, intersections with the Poincaré plane would still be corresponding to the pitch period, less reliable, though. For a noise reduced attractor a singular-value-decomposition (SVD) embedding approach has been proposed [8], but similar results can be achieved by a simple low-pass filter. The latter is computationally less demanding of course, so this is chosen for noise reduction.

Then the signal is upsampled to \( f_s = 96\text{kHz} \) to increase the resolution of the pitch marks, since at low sampling rates the pitch marks would exhibit too much discretisation noise. The embedding in the state space is done by the method of delays, the embedding dimension was chosen be \( M = 9 \). The delay for the chosen sampling frequency is around \( r = 50 \).

---

**Fig. 1. Histogram of space-time separation. The normalized number of state space-distances within a certain neighborhood \( r \) for every time distance \( \Delta t \) is plotted.**

**Fig. 2. State-space embedding before low-pass filtering.**

**Poincaré section**

At the heart of the algorithm is the calculation of the Poincaré hyperplane (fig. 3). Around a chosen point \( x(n_0) \), the neighborhood within a certain radius \( r \) is searched for points. Then a mean flow direction \( f(n_0) \) of the trajectories in this neighborhood \( N(n_0) \) is calculated (considering only those trajectories, with a flow in the same direction as the initial point).

\[
f(n_0) = \text{mean}[x(n + 1) - x(n)] \quad \forall n \in N(n_0)
\]

So for every frame the Poincaré hyperplane is defined as the hyperplane through \( x(n_0) \), which is perpendicular to \( f(n_0) \) (fig. 3).

Mann et al. [8] reported the loss of synchronicity in case of unvoiced portions of the signal. Since in running speech this is usually the case, we decided to use the minimum of the low-pass filtered time-domain signal as an additional criterion for synchronization. So, in every frame we initialize the algorithm with \( n_0 = \min(x) \).

Points in the neighborhood \( N(n_0) \), within a certain distance \( r \) from the plane are considered as pitch mark candidates. Of these candidates, we select those which correspond to an absolute minimum in the time domain.

To remove the influence of a changing amplitude automatic gain control was applied for every frame. This moves the trajectories of quasi-periodic signals closer together, which means, that the attractor is contracted, if it was spread due to amplitude changes.

The length one frame has to be chosen so that at least two periods of the expected minimum frequency fit into the frame. If the signal is periodic, the trajectory returns at least once into the chosen neighborhood and intersects the Poincaré hyperplane and a pitch mark can be detected. The hopsize depends on the the last pitchmark in the current frame. The beginning of the following frame is set to the last pitchmark.

A proper voiced/unvoiced decision is not yet solved. Right now a frame is considered as unvoiced, if no
IV. RESULTS

Formal evaluation of the pitch marking problem still has to be performed. Informal results using the pitch detection evaluation database by Paul Bagshaw [10] (http://www.cstr.ed.ac.uk/projects/fda/) and recordings of dysphonic voices from Graz University Hospital [11] are very promising.

In figure 4 the results of the algorithm on running speech can be seen. The sentence ‘Judith found the manuscripts waiting for her on the piano’ is spoken by a male speaker with modal voice. Most of the pitch marks are correctly set.

The state space plot of the same segment (fig. 6) shows an interesting property. There are two loops with different sizes in the plot. The interpretation is that depending on the period cycle length the state space vector follows either the larger or the smaller loop.

Fig. 7 shows a speech waveform, of a male speaker uttering the German phrase ‘nie und nimmer’. This utterance is described by speech therapists as hoarse, with strong diplophonia and some breathiness; his mean pitch is unusually high for a male person. Besides a few errors the pitch seems to be marked correctly.

Fig. 8 shows a segment of this phrase showing the irregular fundamental period. This case, of course calls for a comparison with a laryngograph signal, which is not available in the database [11].

V. CONCLUSION

An algorithm using Poincaré sections for pitch mark determination for dysphonic voices was presented. The algorithm works on running speech, overcoming the synchronization problem by sticking to the minimum of the time domain signal. A diplophonic case was presented where the alternating pitch period is correctly identified. The results are very promising, and will receive further evaluation.
Fig. 6. Top plot: State-space embedding of the diplophonic speech sample. One can interpret the two loops as the two different attractors for the two fundamental periods. Bottom plot: waveform plot.

Fig. 7. Top plot: Waveform plot of the German phrase 'nie und nimmer'. Bottom plot: fundamental period obtained with Poincare section.

Fig. 8. Top plot: Waveform plot of a segment of the German phrase 'nie und nimmer' and pitch marks. Bottom plot: fundamental period obtained with Poincare section.

REFERENCES


