SUPER-RESOLUTION PITCH ESTIMATOR USING CHAOS THEORY

John Asenstorfer
School of Electrical & Electronic Engineering

University of Adelaide
October 9th 1992

ABSTRACT – Using experimental techniques used in analysing nonlinear dynamical systems a novel pitch estimator is derived. The system allows pitch estimation to a fraction of the sampling period. Issues are addressed that make the estimator reliable and robust.

INTRODUCTION

Obtaining pitch estimators that are inherently stable and accurate has been a problem for many years and as yet remains an unsolved problem. Packet reconstruction for missing packets of speech (at typical packet may be ≈20ms duration) requires an accurate pitch estimator. Traditional pitch determination algorithms use short analysis windows which may contain widely varying speech and voiced and unvoiced segments. The algorithm illustrated here does not rely on such windows although windows are used due to the nature of the packet reconstruction problem.

This work discusses an approach that uses ideas from chaos theory and the experimental methods used in analysing chaotic systems. (Tishby, 1990) used chaos theory to analyse speech.

INITIAL FINDINGS

It has been shown by various authors, for example (Tishby, 1990), that speech shows the properties of a non–linear dynamical system. Generally strange attractors in phase space, associated with chaotic systems, are obtained by time delays of the original time series S(t). A tuplet $[S(t), S(t+\tau), S(t+2\tau), \ldots]$ is created, where the time delay τ is arbitrary for noiseless data but for noisy data it can be shown (Argoul et. al, 1987) that an optimal choice is between one tenth and one half of the mean orbital period. For an overview of the techniques and terminology used in the study of chaotic sytems see for example, (May, 1976), (Argoul et. al, 1987), (Grebogi, Ott & York, 1987).

The work here was interested in the periodic characteristic of the attractor rather than the classification of the strange attractors in speech. This approach differs slightly from the work done previously on chaotic attractors. Another difference is the number of samples available for a particular attractor. Sampling was at 8kHz, and at a pitch of 80Hz held for five pitch periods, gives us 500 samples before the next sound is formed. Chen (Chen, 1988) comments on the amount of data that should be available for numerical algorithms, and comments that for orbits with 10–100 samples about 5–50 orbits are necessary for a correlation dimension D of 2 (Grassberger & Procaccia, 1983). This is of considerable concern as such quantities of data are not available due to the time varying nature of speech. It is noted that Tishby (Tishby, 1990) used 20 segments of voiced speech to obtain an adequate number of samples. His investigations indicate that voiced speech over a range of speakers has an embedding dimension of 3–5. He also notes that unvoiced speech has a higher embedding dimension but his figures may be unreliable due to the limited number of samples available.

Despite the above problems, experiments were undertaken to determine if it was possible to determine the pitch period from phase space. Initial attempts at finding the pitch hinged around the fact that the packets of speech observed in our system were 200 samples long. If the 200 samples are plotted in two dimensions (2D) the periodic nature of the attractor is evident. It was also noted that at the pitch period the orbit tracks came within very close proximity of each other. This suggests that an automated method finding points closest to each other in different orbits should give an estimate of pitch. In what follows the term attractor will be used rather than chaotic attractor.

Poincaré Sections

Poincaré investigated dynamical systems by observing how a phase space track pierced a plane (section in higher dimensions) from a given direction. (Grebogi, Ott & York, 1987) illustrate the idea. The idea of a Poincaré section is useful in our application for finding points close to each other in different cycles of the orbit and extracting pitch information at the same time.

To simplify the discussion we will consider only one plane initially, taken to be the semi–infinite half plane $X \geq 0$. Each time a pair of 3D samples fell on either side of the plane going in an anti–clockwise direction a linear interpolation was made to determine the approximate piercing point of the phase space track made by the system's orbit. A collation is then made of the three pairs of closest piercings, and the number of samples between the pair gives a pitch estimate for each pair.

Naturally something that simple will not suffice in practice, as we can not determine the orientation in phase space of the attractor. To alleviate this problem 3 half planes for the 3D data were used and the resulting data analysed, rejecting cases where the orbit track was on too oblique an angle. This method gave good results for sounds like *i* in wide but became unreliable when the voiced sound changed, producing a range of fractional pitch periods. The fractional pitch periods were characterised by; 1/3, 1/2, 2/3, 3/4, 1, 4/3, 3/2, 2. At times some 1/8 fractions appeared. Although this was interesting it was not followed up because a method that could discriminate the true pitch period was required.

Poincaré Section in Time

Another way of creating a Poincaré section is to try and freeze the position in the orbit using a time section. This is akin to using a stroboscope to freeze the position of a harmonic oscillator by adjusting the strobe time till it closely approximates the period of the system. In this case a mean square error is calculated between points in the present period and that of the previous period for an ensemble of sections (all with the same time lag). The minimum of the averages (over the ensemble), occurs when the best match between the shapes of the attractors is found. This occurs when the best time lapse is chosen and so gives the estimate of the pitch. The expression for this process is given below. This method has similarities to that of Medan (Medan, Yair & Chazan, 1991) although there are significant differences.

$$P_{sr} = \min_{winlo<\lambda<winh} \left\{ \frac{\sum_{i=\varrho}^{\varrho+P_{est}} \left[x(t_i,\tau) - x(t_{i+\lambda},\tau)\right]^2}{P_{est}} \right\}$$

where the value of λ that minimizes the expression is the super resolution pitch estimate P_{sr}. P_{est} is an integral pitch estimate, and winlo and winhi are the lower and upper limits for the range of λ. τ is the delay associated with the construction of the x *tuples*

The technique described above constitutes a considerable computational burden and as such a method for reducing the burden was required. Still staying with the multidimensional phase space notion, an integral–sample pitch estimate was obtained using a multidimensional Absolute Mean Difference Function (amdf). An approximation of the pitch was obtained using amdf to obtain a window of search for the more computationally intensive section search indicated above. It was found that for the 3D and higher dimensional amdf a fairly stable estimate was made. However the amdf made rather characteristic errors and in those cases either the pitch estimate had to be slightly altered and/or the search window for the Poincaré section had to be increased. It became necessary to classify the speech using some technique to adjust the amdf estimate.

SPEECH CLASSIFICATION

The next objective was to construct a simple speech classification system that was pitch and speaker independent. A simple technique that has been used in the past is zero crossing. Zero crossing however did not give all the information that was required and some ideas from fractal theory were investigated. Fractal theory gives ways of classifying "rough" contours. In fractal theory, the set of zero crossings is called the zero set. The zero set however gives no information about amplitude between zero crossings, so another method was developed, the "Devil's Staircase". For details of the more traditional "Devils Staircase" see (Feder, 1988).

The Zero Set

To obtain a classification for voiced and unvoiced speech it is quite evident that the rate of zero crossings is important. Unvoiced speech is characterized by a very high zero crossing rate. I used a very simple mechanism based on the number of samples between consecutive crossings to obtain two parameters; fine structure and coarse structure. After experimentation it was determined that fine structure be classed as that which has less than four samples between consecutive crossings (sampling at 8kHz) and if greater it would be classed as coarse. It was desirable to have a continuous "real number" to classify "coarse" and "fine" to simplfy the overall pitch estimation algorithm and avoid counting over windowed speech. A recursive form used to estimate the "fine" parameter, is given by;

$$x_{n+1} = y + (x_n)^\alpha$$

$$y = \begin{cases} 1 & \text{if no. of samples between this crossing and last} < 4 \\ 0 & \text{if number of samples} \geq 4 \end{cases}$$

where the equation is evaluated at every zero crossing n and x_{n+1} is the value of the parameter at the (n+1)th zero crossing. The power α determines the "memory" of the parameter by regulating the decay. If α is too great the response of the parameter to the actual conditions is too slow. A value of 0.85 was found to be useful for the current application. This method gives a geometric series and the range of values can be readily determined.

The parameter for "coarse" was defined analogously. Although the contour of the parameter is rough it proved valuable, see Fig. 1. For a sibilant like *s* the "fine" parameter takes on values around 9.

The Devil's Staircase

The Devil's Staircase was used primarily to find information about amplitude but gave extra information as well. The staircase as constructed here is not the same as the traditional staircase. It was constructed for this application using

$$M_I = \sum_{i}^{\infty} \Phi_i$$

$$\Phi_i = \begin{cases} max\{|s_n|; z_{i-1} < |s_n| \leq z_i\} & \text{if a zero crossing occurred} \\ 0 & \text{otherwise} \end{cases}$$

where to give the staircase, the maximum of the absolute value of speech samples s_n between the zero crossings z_{i-1} and z_i is summed when a zero crossing occurred otherwise 0 is added.

It was found that for different speakers recorded under similar conditions the average slope of the staircase was similar and that the slope for voiced and unvoiced speech was also similar as shown in Fig 1. Where there is no speech the slope is almost zero and where voicing changes within a word such as "wagon", the slope decreases in the region of change.

THE PITCH ESTIMATION ALGORITHM

The pitch estimation algorithm consists of the following units

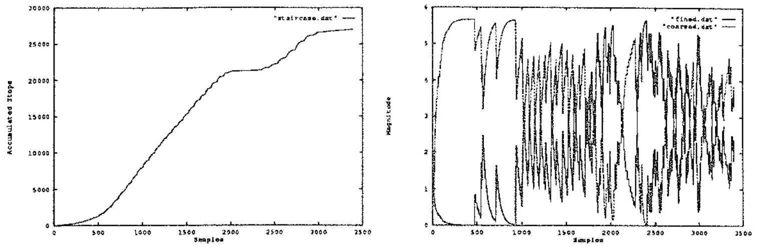

Figure 1: The Devil's Staircase and structure parameters for "wagon"

Parameters — Determine the "fine" and "coarse" speech structure parameters and also the average slope of the staircase. These are used with a few heuristic rules for adjusting the estimate from the amdf search.

amdf Search — An amdf search is undertaken in the number of dimensions chosen and returns an integral sample estimate of the pitch period. Often the estimate will be within a sample or two of the final pitch estimate, but not always. The search window can be reduced by making use of the fact that the pitch variation almost never exceeds 25%, (Medan, Yair & Chazan, 1991).

Section Search — The estimate from the amdf search is used to give a window to search for the non-integral pitch estimate. A golden-mean section search was used to find the estimate although a gradient technique could be used especially for higher dimensions. The well known Lagrange interpolation formula (second order) was used to interpolate sample values.

Feed–back — The pitch estimate is fed back to assist pitch tracking. It is also used to adjust delay in obtaining the multidimensional phase space representation.

The *parameters* module and *amdf* module logically occur in parallel to deliver an adjusted estimate of the integral pitch and search window to the *section search*. The final estimate is fed back to assist in pitch tracking. A set of simple heuristic rules using the parameters is used in pitch tracking. An example of one such rule is that if a word starts (the slope on the staircase increases beyond a threshold) with a fricative (high value for the fine structure parameter and a low value for the coarse structure parameter) the *amdf* tends to estimate too high for the pitch. At the point where voicing starts (coarse structure parameter's value increases) the estimate is still too high. In this case the estimated pitch is decreased by 80% but the window for the *section search* is increased to include the original *amdf* estimate.

Another important element in pitch tracking is that the new integral pitch estimate must be within 25% of the last pitch estimate. If it is not, the old pitch estimate is used and the *section search* has to find the best pitch.

RESULTS

Testing a pitch estimator of this kind without using synthetic speech is difficult. I chose to use different words from four speakers, two male and two female, to test the estimator. Hand measurements of the pitch throughout each word were made as a guide to judge if the pitch estimator was operating correctly.

It was found that the pitch estimate varied a little depending on the actual delay τ chosen for the multidimensional phase space representation. To stabilize the measurements it was found that a value

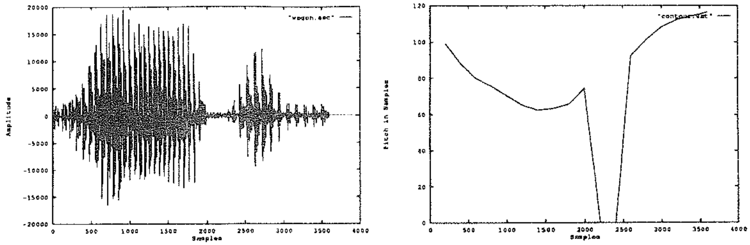

Figure 2: The speech waveform and pitch contour for "wagon"

of approximately 1/9 of the *actual pitch* gave good results. Generally a value larger than the threshold of around a tenth of the pitch period gave fairly stable results, but using a value that is too large tended to produce a "smearing" effect where too many pitch perods are taken into consideration.

The estimator was trialled using up to six phase space dimensions. The results indicate that using more than three dimensions is not justified. The amdf function proved to be a little more reliable in three dimensions than lower dimensions but showed no change above three. The section search showed little change above three dimensions and there is little point in going beyond this. Using three dimensions and finding the mean square distance over the whole pitch constitutes an averaging (low–pass filtering), and as such no pre-lowpass filtering was deemed necessary. Tests were performed for three dimensions with and without lowpass filtered data files, and no significant difference was found. The higher dimensions provided greater averaging and the distance function was smoother, giving slightly different values to lower dimensional estimates. However "smearing" between three or more pitch periods occurs for the higher dimensions, particularly if the chosen delays are large.

word	mean standard dev. for dimensions 1–6 (samples 8kHz)	mean standard dev. for dimensions 3–6 (samples 8kHz)
wagon	6.23	0.18
purse	0.42	0.18
droopy	0.54	0.04
fine	0.21	0.02

Table 1. Expected standard deviations on pitch estimate for different dimension ranges

To further increase the robustness (to avoid the 7/8 pitch frequency sometimes present) a variation was used in the distance function. Four different phases and lengths of the pitch period were used in calculating the distance. Each phase is started at a fixed place in the pitch period and only every fourth multidimensional point taken in the distance calculation. The four phases are staggered so that most of the multidimensional points in a pitch period are accounted for. The four distances are then summed and averaged. The true pitch is characterized by a high correlation for any segment within consecutive pitch periods. This argument extends to different length segments, as a length a little longer than a pitch period should still find the best match, in a mean square error sense, one pitch period back.

Figure 2 shows the speech wave form for the word "wagon" spoken by a male speaker with the pitch contour. Note the pitch estimates were form only for the last pitch in a 200 sample window, as that is what is required for the packet reconstruction problem.

Various methods of measuring closeness were also investigated; mean square error, mean absolute error and correlation. All methods yielded similar but not exactly the same results. The pitch estimate is a statistical estimate and as such an accuracy of a tenth of a sampling period seemed reasonable. Medan (Medan, Yair & Chazan, 1991) imply that greater resolution is possible in their technique, but this does not seem possible. For an accuracy of a tenth of a sampling period an average of approximately seven iterations were required in the Poincaré section search.

SUMMARY

A reliable and robust super resolution pitch estimator, which performs well, was discussed and will give an accuracy to 1/10 of a sample (sampling at 8kHz). As no preprocessing is required and the multidimensional phase space representation is obtained by simple delays, the computational load is not great although greater reductions in processing would be possible. It was interesting to note that a large number of dimensions in representing the speech did not significantly improve performance. One of the greatest gains was to use three dimensions for the amdf to improve its reliability and estimate.

REFERENCES

Argoul, F. et al (1987) "Chemical Chaos: From Hints to Confirmation", Accounts Of Chem. Research, Vol 20, Dec.

Chen, P. (1988) "Empirical and Theoretical Evidence of Economic Chaos", Systems Dynamics Review Vol. 4 Numbers 1–2

Feder,J. (1988) *Fractals*, pp67-73, (Plenum Press)

Grassberger, P. & I.Procaccia (1983) "Measuring the Strangeness of Strange Attractors", Physica D 9:189-192

Grebogi,C., E.Ott & J.A.Yorke (1987) "Chaos, Strange Attractors, and Fractal Basin Boundaries in Nonlinear Dynamics", Science, Vol 238, Oct.

May,R.M. (1976) "Simple Mathematical Models With Very Complicated Dynamics", Nature Vol. 261 June 10

Medan,Y., E.Yair, & D.Chazan (1991) "Super Resolution Pitch Determination of Speech Signals", IEEE Trans. on Signal Proc. Vol. 39, No. 1. Jan.

Tishby, N. (1990) "A Dynamical Systems Approach to Speech Processing", S6b.5 ICASSP 90