5 Nonlinear methods for speech analysis and synthesis

Steve McLaughlin and Petros Maragos

5.1. Introduction

Perhaps the first question to ask on reading this chapter is why should we consider nonlinear methods as offering any insight into speech signals given the success of current, mostly linear-based, speech analysis methods. There are known to be a number of nonlinear effects in the speech production process. Firstly, it has been accepted for some time that the vocal tract and the vocal folds do not function independently of each other, but that there is in fact some form of coupling between them when the glottis is open [27] resulting in significant changes in formant characteristics between open and closed glottis cycles [7]. More controversially, Teager and Teager [69] have claimed (based on physical measurements) that voiced sounds are characterized by highly complex air flows in the vocal tract involving jets and vortices, rather than well-behaved laminar flow. In addition, the vocal folds will themselves be responsible for further nonlinear behavior, since the muscle and cartilage which comprise the larynx have nonlinear stretching qualities. Such nonlinearities are routinely included in attempts to model the physical process of vocal fold vibration, which have focused on two or more mass models [24, 28, 65], in which the movement of the vocal folds is modeled by masses connected by springs, with nonlinear coupling. Observations of the glottal waveform have shown that this waveform can change shape at different amplitudes [59] which would not be possible in a strictly linear system where the waveform shape is unaffected by amplitude changes. Models of the glottal pulse also include nonlinearities, for example, the use of nonlinear shaping functions [58–60] or the inclusion of nonlinear flow [22].

In order to arrive at the simplified linear model, a number of major assumptions are made:

(i) the vocal tract and speech source are uncoupled (thus allowing source-filter separation);

(ii) airflow through the vocal tract is laminar;

(iii) the vocal folds vibrate in an exactly periodic manner during voiced speech production;

(iv) the configuration of the vocal tract will only change slowly.

These imply a loss of information which means that the full speech signal dynamics can never be properly captured. These inadequacies can be seen in practice in speech synthesis where, at the waveform generation level, current systems tend to produce an output signal that lacks naturalness. This is true even for concatenation techniques which copy and modify actual speech segments.

Given the statements above then should make clear why nonlinear methods will offer useful insights and suggest useful methods that we can adopt to enhance speech synthesis. The chapter is structured as follows. First some discussion on speech aerodynamics and modulations in speech are discussed. Then the discussion moves on to consider why conventional linear methods work as well as they do. The discussion then moves on to consider what nonlinear methods we might use and for what. These range from modulation models, fractal methods, using Poincaré maps for epoch detection, the use of unstable manifolds, and functional approximation methods. Then consideration is briefly given to the use of nonlinear methods in automatic speech recognition. Finally some conclusions are drawn and suggestions for potential areas of research are considered.

5.1.1. Speech aerodynamics

For several decades the traditional approach to speech modeling has been the linear (source-filter) model where the true nonlinear physics of speech production is approximated via the standard assumptions of linear acoustics and 1D plane wave propagation of the sound in the vocal tract. This approximation leads to the well-known linear prediction model for the vocal tract where the speech formant resonances are identified with the poles of the vocal tract transfer function. The linear model has been applied to speech coding, synthesis, and recognition with limited success [55, 56]. To build successful applications, deviations from the linear model are often modeled as second-order effects or error terms. However, there is strong theoretical and experimental evidence [2, 25, 36, 40, 62, 69, 70] for the existence of important nonlinear aerodynamic phenomena during the speech production that cannot be accounted for by the linear model. Thus, the linear model can be viewed only as a first-order approximation to the true speech acoustics which also contain second-order and nonlinear structure. The investigation of speech nonlinearities can proceed in at least two directions: (i) numerical simulations of the nonlinear differential (Navier-Stokes) equations [72] governing the 3D dynamics of the speech airflow in the vocal tract, as, for example, in [57, 70], and (ii) development of nonlinear signal processing systems suitable to detect such phenomena and extract related information. Most of the research in this aspect of nonlinear speech processing, for example, as reviewed in [35, 54], has focused on the second approach, which is computationally much simpler, that is, to develop models and extract related acoustic signal features describing two types of nonlinear phenomena in speech, *modulations* and *turbulence*. Turbulence can be explored both from the geometric aspect, which brings us to *fractals* [32], and from the nonlinear dynamics aspect, which leads us to *chaos* [1, 47].

In this chapter we summarize the main concepts, models, and algorithms that have been used or developed in the three above nonlinear methodologies for speech analysis and synthesis.

5.1.2. Speech turbulence

Conservation of momentum in the air flow during speech production yields the Navier-Stokes equation [72]:

$$\rho\left(\frac{\partial \vec{u}}{\partial t}+\vec{u}\cdot\nabla\vec{u}\right)=-\nabla p+\mu\nabla^2\vec{u}, \tag{5.1}$$

where ρ is the air density, p is the air pressure, $\vec{u}$ is the (vector) air particle velocity, and μ is the air viscosity coefficient. It is assumed that flow compressibility is negligible [valid since in speech flow (Mach numbers)$^2 \ll 1$] and hence $\nabla\cdot\vec{u}=0$. An important parameter characterizing the type of flow is the Reynolds number $\mathrm{Re}=\rho UL/\mu$, where U is a velocity scale for $\vec{u}$ and L is a typical length scale, for example, the tract diameter. For the air we have very low μ and hence high Re. This causes the inertia forces to have a much larger order of magnitude than the viscous forces $\mu\nabla^2\vec{u}$. A *vortex* is a region of similar (or constant) vorticity $\vec{\omega}$, where $\vec{\omega}=\nabla\times\vec{u}$. Vortices in the air flow have been experimentally found above the glottis by Teager and Teager [69] and Thomas [70] and theoretically predicted by Kaiser [25], Teager and Teager [69], and McGowan [40] using simple geometries. There are several mechanisms for the creation of vortices: (1) velocity gradients in boundary layers, (2) separation of flow, which can easily happen at cavity inlets due to adverse pressure gradients (see [69] for experimental evidence), and (3) curved geometry of tract boundaries, where due to the dominant inertia forces the flow follows the curvature and develops rotational components. After a vortex has been created, it can propagate downstream as governed by the vorticity equation [72]

$$\frac{\partial\vec{\omega}}{\partial t}+\vec{u}\cdot\nabla\vec{\omega}=\vec{\omega}\cdot\nabla\vec{u}+\nu\nabla^2\vec{\omega}, \quad \nu=\frac{\mu}{\rho}. \tag{5.2}$$

The term $\vec{\omega}\cdot\nabla\vec{u}$ causes vortex twisting and stretching, whereas $\nu\nabla^2\vec{\omega}$ produces diffusion of vorticity. As Re increases (e.g., in fricative sounds or during loud speech), all these phenomena may lead to instabilities and eventually result in *turbulent flow*, which is a “state of continuous instability” [72] characterized by broad-spectrum rapidly varying (in space and time) velocity and vorticity. Many speech sounds, especially fricatives and stops, contain various amounts of turbulence. In the linear speech model this has been dealt with by having a white noise source exciting the vocal tract filter.

Modern theories that attempt to explain turbulence [72] predict the existence of eddies (vortices with a characteristic size λ) at multiple scales. According to the energy cascade theory, energy produced by eddies with large size λ is transferred hierarchically to the small-size eddies which actually dissipate this energy due to viscosity. A related result is the Kolmogorov law

$$E(k,r) \propto r^{2/3}k^{-5/3}, \tag{5.3}$$

where $k = 2\pi/\lambda$ is the wavenumber in a finite nonzero range, r is the energy dissipation rate, and $E(k,r)$ is the velocity wavenumber spectrum, that is, Fourier transform of spatial correlations. This multiscale structure of turbulence can in some cases be quantified by *fractals*. Mandelbrot [32] and others have conjectured that several geometrical aspects of turbulence (e.g., shapes of turbulent spots, boundaries of some vortex types found in turbulent flows, shape of particle paths) are fractal in nature. We may also attempt to understand aspects of turbulence as cases of chaos. Specifically, chaotic dynamical systems converge to attractors whose sets in phase space or related time series signals can be modeled by fractals; references can be found in [47]. Now there are several mechanisms in high-Re speech flows that can be viewed as routes to chaos; for example, vortices twist, stretch, and fold (due to the bounded tract geometry) [32, 72]. This process of twisting, stretching, and folding has been found in low-order nonlinear dynamical systems to give rise to chaos and fractal attractors.

5.1.3. Speech modulations

By "speech resonances" we will loosely refer to the oscillator systems formed by local vocal tract cavities emphasizing certain frequencies and de-emphasizing others. Although the linear model assumes that each speech resonance signal is a damped cosine with constant frequency within 10–30 ms and exponentially decaying amplitude, there is much experimental and theoretical evidence for the existence of *amplitude modulation* (AM) *and frequency modulation* (FM) in speech resonance signals, which make the amplitude and frequency of the resonance vary instantaneously within a pitch period. First, due to the *airflow separation* [69], the air jet flowing through the vocal tract during speech production is highly unstable and oscillates between its walls, attaching or detaching itself, and thereby changing the effective cross-sectional areas and air masses. This can cause modulations of the air pressure and velocity fields, because slow time variations of the elements of simple oscillators can result in amplitude or frequency modulation of the oscillator's sinusoidal response. Also, during speech production vortices can easily be generated and propagated along the vocal tract [70, 72], while acting as modulators of the energy of the jet. Motivated by this evidence, Maragos et al. [36] proposed to model each speech resonance with an AM-FM signal

$$x(t) = a(t)\cos[\phi(t)] = a(t)\cos\left[2\pi\int_0^t f(\tau)d\tau\right] \tag{5.4}$$

and the total speech signal as a superposition of such AM-FM signals, $\sum_k a_k(t) \cos[\phi_k(t)]$, one for each formant. Here $a(t)$ is the instantaneous amplitude signal and $f(t)$ is the instantaneous cyclic frequency representing the time-varying formant signal. The short-time formant frequency average $f_c = (1/T)\int_0^T f(t)dt$, where T is in the order of a pitch period, is viewed as the carrier frequency of the AM-FM signal. The classical linear model of speech views a formant frequency as constant, that is, equal to f_c, over a short-time (10–30 ms) frame. However, the AM-FM model can both yield the average f_c and provide additional information about the formant's instantaneous frequency deviation $f(t) - f_c$ and its amplitude intensity $|a(t)|$.

5.1.4. So if speech is nonlinear, why do linear methods work? Conventional speech synthesis approaches

Conventionally the main approaches to speech synthesis depend on the type of modeling used. This may be a model of the speech organs themselves (articulatory synthesis), a model derived from the speech signal (waveform synthesis), or alternatively the use of prerecorded segments extracted from a database and joined together (concatenative synthesis).

Modeling the actual speech organs is an attractive approach, since it can be regarded as being a model of the fundamental level of speech production. An accurate articulatory model would allow all types of speech to be synthesized in a natural manner, without having to make many of the assumptions required by other techniques (such as attempting to separate the source and vocal tract parts out from one signal) [19, 24, 28]. Realistic articulatory synthesis is an extremely complex process, and the data required is not at all easy to collect. As such, it has not to date found any commercial application and is still more of a research tool.

Waveform synthesizers derive a model from the speech signal as opposed to the speech organs. This approach is derived from the linear source-filter theory of speech production [17]. The simplest form of waveform synthesis is based on linear prediction (LP) [38]. The resulting quality is extremely poor for voiced speech, sounding very robotic.

Formant synthesis uses a bank of filters, each of which represents the contribution of one of the formants. The best known formant synthesizer is the Klatt synthesizer [26], which has been exploited commercially as DECTalk. The synthesized speech quality is considerably better than that of the LP method, but still lacks naturalness, even when an advanced voice-source model is used [16].

Concatenation methods involve joining together prerecorded units of speech which are extracted from a database. It must also be possible to change the prosody of the units, so as to impose the prosody required for the phrase that is being generated. The concatenation technique provides the best quality synthesized speech available at present. It is used in a large number of commercial systems, including British Telecom's Laureate [45] and the AT&T Next-Gen system [3]. Although there is a good degree of naturalness in the synthesized output, it is still clearly distinguishable from real human speech, and it may be that more sophisticated parametric models will eventually overtake it.

Techniques for time and pitch scaling of sounds held in a database are also extremely important. Two main techniques for time-scale and pitch modification in concatenative synthesis can be identified, each of which operates on the speech signal in a different manner. The pitch synchronous overlap add (PSOLA) [41] approach is nonparametric as opposed to the harmonic method, which actually decomposes the signal into explicit source and vocal tract models. PSOLA is reported to give good quality, natural-sounding synthetic speech for moderate pitch and time modifications. Slowing down the speech by a large factor (greater than two) does introduce artifacts due to the repetition of PSOLA bells. Some tonal artifacts (e.g., whistling) also appear with large pitch scaling, especially for higher pitch voices, such as female speakers and children.

McAulay and Quatieri developed a speech generation model that is based on a glottal excitation signal made up of a sum of sine waves [39]. They then used this model to perform time-scale and pitch modification. Starting with the assumption made in the linear model of speech that the speech waveform $x(t)$ is the output generated by passing an excitation waveform $e(t)$ through a linear filter $h(t)$, the excitation is defined as a sum of sine waves of arbitrary amplitudes, frequencies, and phases. A limitation of all these techniques is that they use the linear model of speech as a basis.

5.2. What nonlinear methods might we use?

5.2.1. Modulation model and energy demodulation algorithms

In the modulation speech model, each speech resonance is modeled as an AM-FM signal and the total speech signal as a superposition of several such AM-FM signals, one for each formant. To isolate a single resonance from the original speech signal, bandpass filtering is first applied around estimates of formant center frequencies.

For demodulating a single resonance signal, Maragos et al. [36] used the nonlinear Teager-Kaiser energy-tracking operator $\Psi[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t)$, where $\dot{x} = dx/dt$, to develop the following nonlinear algorithm:

$$\sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}} \approx 2\pi f(t), \qquad \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}} \approx |a(t)|. \tag{5.5}$$

This is the *energy separation algorithm* (ESA) and it provides AM-FM demodulation by tracking the physical energy implicit in the source producing the observed acoustic resonance signal and separating it into its amplitude and frequency components. It yields very good estimates of the instantaneous frequency signal $f(t) \geq 0$ and of the amplitude envelope $|a(t)|$ of an AM-FM signal, assuming that $a(t)$, $f(t)$ do not vary too fast (small bandwidths) or too greatly compared with the carrier frequency f_c.

There is also a *discrete* version of the ESA, called DESA [36], which is obtained by using a discrete energy operator on discrete-time nonstationary sinusoids. The DESA is an efficient approach to demodulating speech resonances for the following several reasons. (i) It yields very *small errors* for AM-FM demodulation. (ii) It has an extremely *low computational complexity*. (iii) It has an excellent time resolution, almost *instantaneous*; that is, operates on a 5-sample moving window. (iv) It has a useful and intuitive interpretation of tracking and separating the true physical energy of the acoustic source. (v) It can detect transient events. Extensive experiments on speech demodulation using the DESA in [36, 51, 52] indicate that these amplitude/frequency modulations *exist* in real speech resonances and are necessary for its *naturalness*, as found from synthesizing speech via an AM-FM vocoder [52] that uses the AM-FM model.

The main disadvantage of the DESA is a moderate sensitivity to noise. This can be reduced by using *regularized* versions of the continuous ESA adapted for discrete data. Two such continuous approaches were developed by Dimitriadis and Maragos [13]. The first, called *Spline*-ESA, interpolates the discrete-time signal with *smoothing splines* to create a continuous-time signal, applies the continuous-time ESA (5.5), and finally samples the information-bearing signals to obtain estimates of the instantaneous amplitude and frequency of the original discrete signal. In the second approach, called *Gabor*-ESA, the signal derivatives in the original ESA are replaced by signal convolutions with corresponding derivatives of the Gabor filters' impulse response.

The ESAs are efficient demodulation algorithms only when they are used on narrowband AM-FM signals [6]. This constraint makes the use of *filterbanks* (i.e., parallel arrays of bandpass filters) inevitable for wideband signals like speech. Thus, each short-time segment (analysis frame) of a speech signal is simultaneously filtered by all the bandpass filters of the filterbank, and then each filter output is demodulated using the ESA. Ongoing research in speech modulations has been using filterbanks with Gabor bandpass filters whose center frequencies are spaced either linearly or on a mel-frequency scale [13, 52]. Figure 5.1 shows an example of demodulating three bands of a speech phoneme into their instantaneous amplitude and frequency signals.

While the instant frequency signals produced by demodulating resonances of speech vowels have a quasiperiodic structure, those of fricatives look random. Since fricative and stop sounds contain turbulence, Maragos and Dimakis [11, 35] proposed a *random modulation* model for resonances of fricatives and stops where the instant phase modulation signal is a random process from the $1/f$ noise family. Specifically, they modeled each such speech resonance $R(t)$ as

$$R(t) = a(t)\cos\left(2\pi f_c t + p(t)\right), \qquad E\Big[\,|P(\omega)|^2\Big] \propto \frac{\sigma^2}{|\omega|^\gamma}, \tag{5.6}$$

where $p(t)$ is a random nonlinear phase signal, $P(\omega)$ is its Fourier transform, and $E[\cdot]$ denotes expectation. The power spectral density (PSD), measured either by a sample periodogram $|P(\omega)|^2$ or empirically via filterbanks, is assumed to obey a $1/\omega^\gamma$ power law; such processes are called the "$1/f$ noises." In [35] the

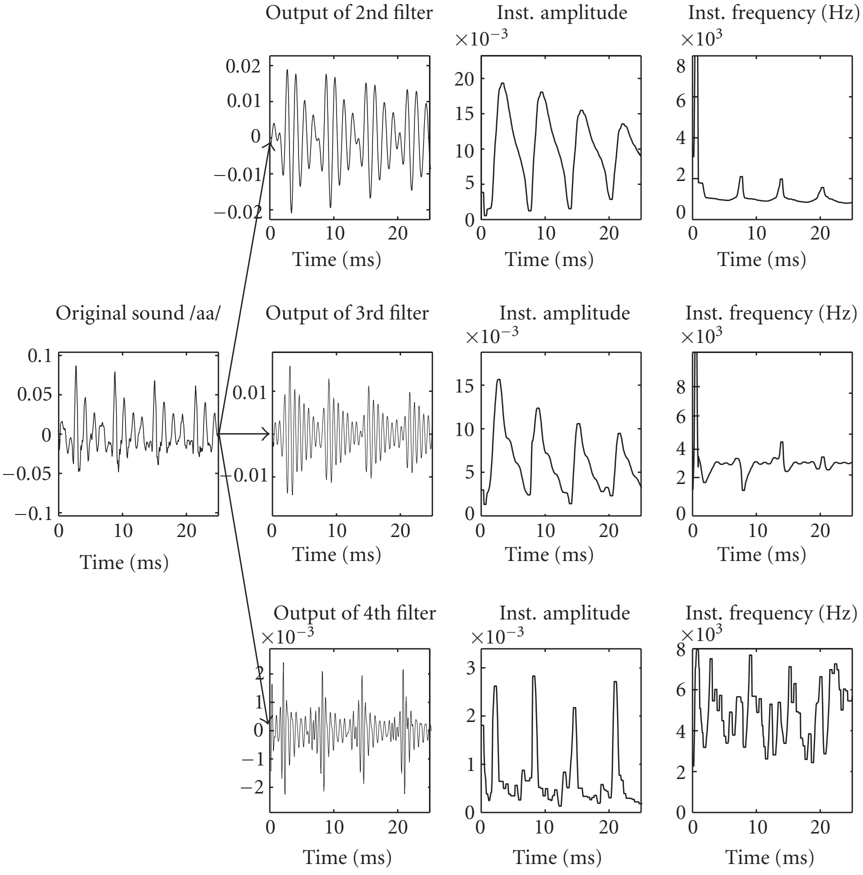


FIGURE 5.1. Demodulating a speech phoneme using a Gabor filterbank and the Spline-ESA.

proposed $1/f$ model for the instant phase was the fractional Brownian motions (FBMs), which are a popular fractal model for a subclass of $1/f$ noises [32]. In [11] this $1/f$ phase model was extended to include the class of *alpha-stable* processes. The method used in [11, 35] to solve the inverse problem, that is, that of extracting the phase modulation $p(t)$ from the speech resonance and modeling it as a $1/f$ noise, is summarized in the following steps.

(1) Isolate the resonance by Gabor bandpass filtering the speech signal.
(2) Use the ESA to estimate the AM and FM signals, $a(t)$ and $f(t)$.
(3) Median filter the FM signal for reducing some extreme spikes.
(4) Estimate the phase modulation signal $p(t)$ by integrating the instant frequency.
(5) Estimate the spectral exponent γ of the phase modulation signal by using a statistically robust estimator of its power spectrum and least-squares fitting a line only on the part of the power spectrum not affected by lowpass filtering.

The efficiency of this method was successfully tested on artificial resonance signals with $1/f$ phase modulation signals.

Strong experimental evidence was also presented that certain classes of speech signals have resonances that can be effectively modeled as phase modulated $1/f$ signals. The validity of the model was demonstrated by confirming that its power spectrum obeys a spectral $1/f^{\gamma}$ power law.

Figure 5.2 demonstrates the application of the above described $1/f$ phase modulation model to a voiced fricative. In [11, 35] extensive similar experiments have been performed on real speech signals (from the TIMIT database), by following the same procedure: a strong speech resonance is located, possibly by using the iterative ESA method. Then the ESA is used to extract the phase modulation. (The phase modulations were also estimated via the Hilbert transform to make sure that the ESA does not introduce any artifacts.) The estimated phase is assumed to be a low-passed version of a $1/f^{\gamma}$ random process and the γ exponent is estimated from the slope of the power spectrum. In all these experiments the conjecture in [35] that the phase modulation of random speech resonances has a $1/f^{\gamma}$ spectrum has always been verified.

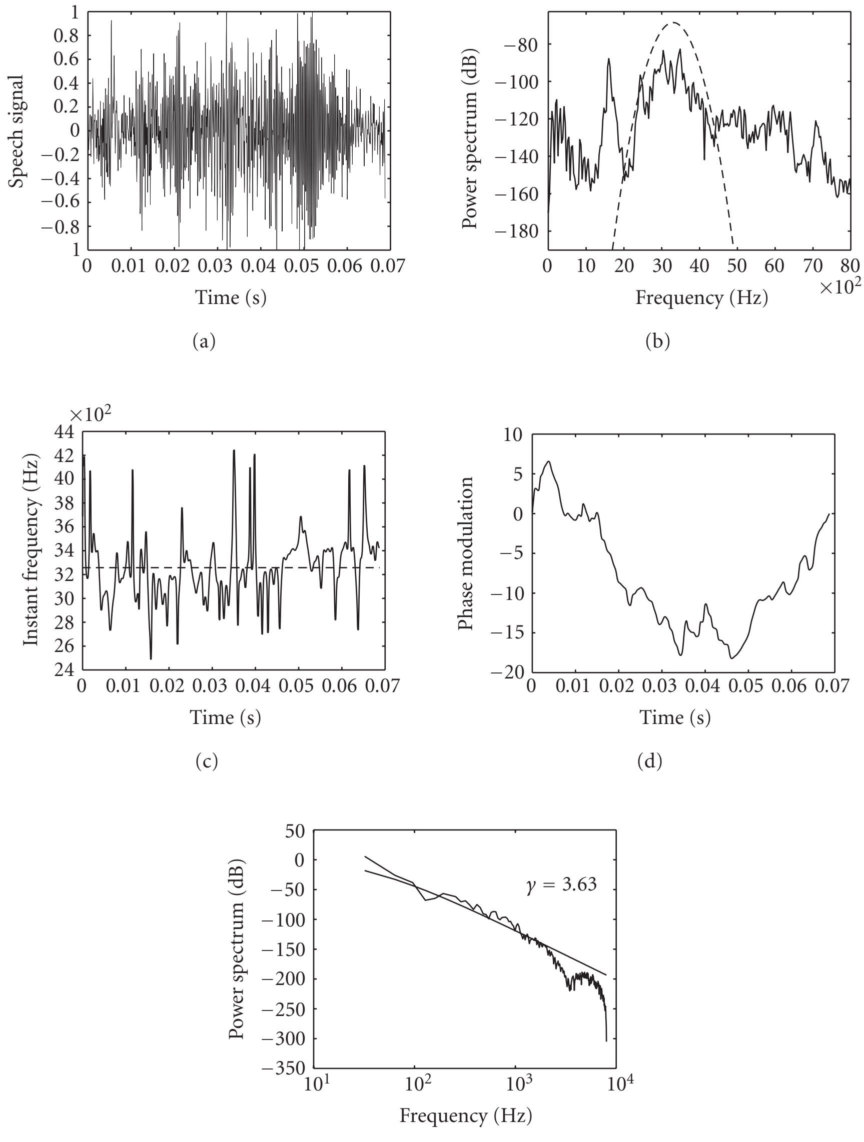
Ongoing work in this area includes better estimation algorithms and a statistical study relating estimated exponents with types of sounds. Some advances can be found in [11].

5.2.2. Fractal methods

Motivated by Mandelbrot's conjecture that fractals can model multiscale structures in turbulence, Maragos [34] used the *short-time fractal dimension* of speech sounds as a feature to approximately quantify the degree of turbulence in them. Although this may be a somewhat simplistic analogy, the short-time fractal dimension of speech has been found in [34, 37] to be a feature useful for speech sound classification into phonetic classes, segmentation, and recognition. An efficient algorithm developed in [34] to measure it consists of using multiscale morphological filters that create geometrical covers around the graph of the speech signal, whose fractal dimension D can then be found by

$$D = \lim_{s \to 0} \frac{\log\left[\text{Area of dilated graph by disks of radius } s/s^{2}\right]}{\log(1/s)}, \qquad (5.7)$$

D is between 1 and 2 for speech signals; the larger D is, the larger the amount of geometrical fragmentation of the signal graph is. In practice, real-world signals do not have the same structure over all scales; hence, D is computed by least-squares fitting a line to the log-log data of (5.7) over a small scale window that can move along the s axis and thus create a profile of local *multiscale fractal dimensions* $D(s,t)$ at each time location t of the short speech analysis frame. The function $D(s,t)$ is called a *fractogram*. The fractal dimension at the smallest scale ($s = 1$) can provide some discrimination among various classes of sounds such as vowels (very low D), unvoiced fricatives (very high D), and voiced fricatives (medium D). At higher



(a)

(b)

(c)

(d)

(e)

Figure 5.2. Experiments with phoneme /z/. (a) Speech signal $s(t)$, (b) PSD of $s(t)$ and Gabor filter, (c) instant frequency, (d) phase modulation $\hat{p}(t)$, and (e) PSD of $\hat{p}(t)$ and estimated slope.

scales, the fractogram multiscale fractal dimension profile can also offer additional information that helps in discriminating among speech sounds, see Figure 5.3.

Related to the Kolmogorov 5/3 law (5.3) is the fact that the variance between particle velocities at two spatial locations X and $X + \Delta X$ varies $\propto |\Delta X|^{2/3}$. By

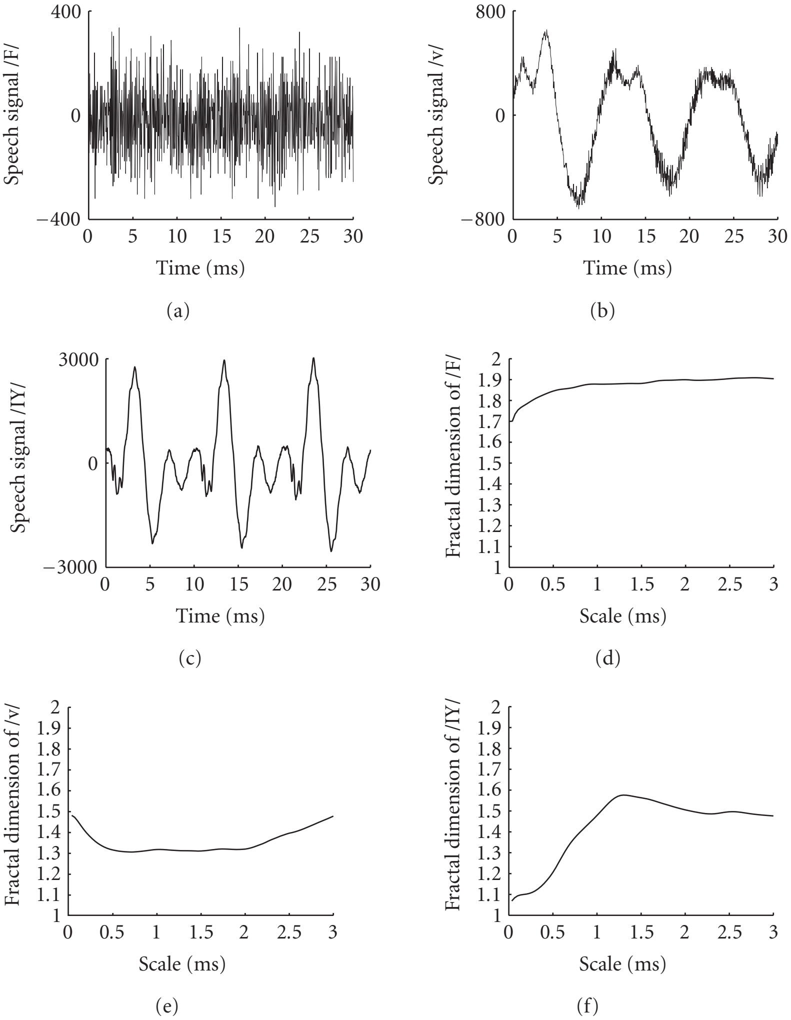


Figure 5.3. (a), (b), and (c) show waveforms from speech sounds sampled at 30 kHz. (d), (e), and (f) show their multiscale fractal dimensions estimated over moving windows of 10 scales.

linking this to similar scaling laws in FBMs, it was concluded in [34] that speech turbulence leads to fractal dimension of $D = 5/3$, which was often approximately observed during experiments with fricatives.

5.2.3. Poincaré maps and epoch marking

The section discusses how nonlinear techniques can be applied to pitch marking of continuous speech. We wish to locate the instants in the time domain speech signal at which the glottis is closed. A variety of existing methods can be employed to locate the epochs. These are abrupt change detection [10], maximum likelihood epoch detection [9], and dynamic programming [68]. All of the above techniques are sound and generally provide good epoch detection. The technique presented here should not be viewed as a direct competitor to the methods outlined above. Rather it is an attempt to show the practical application of ideas from nonlinear dynamical theory to a real speech processing problem. The performance in clean speech is comparable to many of the techniques discussed above.

In nonlinear processing a d-dimensional system can be reconstructed in an m-dimensional state space from a single dimension time series by a process called embedding. Takens' theorem states that $m \geq 2d+1$ for an adequate reconstruction [67], although in practice it is often possible to reduce m. An alternative is the singular value decomposition (SVD) embedding [8], which may be more attractive in real systems where noise is an issue.

A Poincaré map is often used in the analysis of dynamical systems. It replaces the flow of an nth order continuous system with an $(n-1)$th order discrete-time map. Considering a three-dimensional attractor a Poincaré section slices through the flow of trajectories and the resulting crossings form the Poincaré map. Reexamining the attractor reconstructions of voiced speech shown above, it is evident that these three-dimensional attractors can also be reduced to two-dimensional maps.[1] Additionally, these reconstructions are pitch synchronous, in that one revolution of the attractor is equivalent to one pitch period. This has previously been used for cyclostationary analysis and synchronization [31]; here we examine its use for epoch marking.

The basic processing steps required for a waveform of N points are as follows.

(1) Mark y_{GCI}, a known glottal closure instant (GCI) in the signal.

(2) Perform an SVD embedding on the signal to generate the attractor reconstruction in 3D state space.

(3) Calculate the flow vector, $\mathbf{h}$, at the marked point $\mathbf{y}_{\mathrm{GCI}}$ on the attractor.

(4) Detect crossings of the Poincaré section, Σ, at this point in state space by signs changes of the scalar product between $\mathbf{h}$ and the vector $\mathbf{y}_i - \mathbf{y}_{\mathrm{GCI}}$ for all $1 \leq i \leq N$ points.

(5) Points on Σ which are within the same portion of the manifold as $\mathbf{y}_{\mathrm{GCI}}$ are the epochs.

When dealing with real speech signals a number of practical issues have to be considered. The input signal must be treated on a frame-by-frame basis, within which the speech is assumed stationary. Finding the correct intersection points on the Poincaré section is also a difficult task due to the complicated structure of the attractor. Because of this, additional measures are used for locating the epoch points. The flow chart shown in Figure 5.4 illustrates the entire process. Two different data sets were used to test the performance of the algorithm, giving varying degrees of realistic speech and hence difficulty.

[1] Strictly these attractor reconstructions are discrete-time maps and not continuous flows. However it is possible to construct a flow vector between points and use this for the Poincaré section calculation.

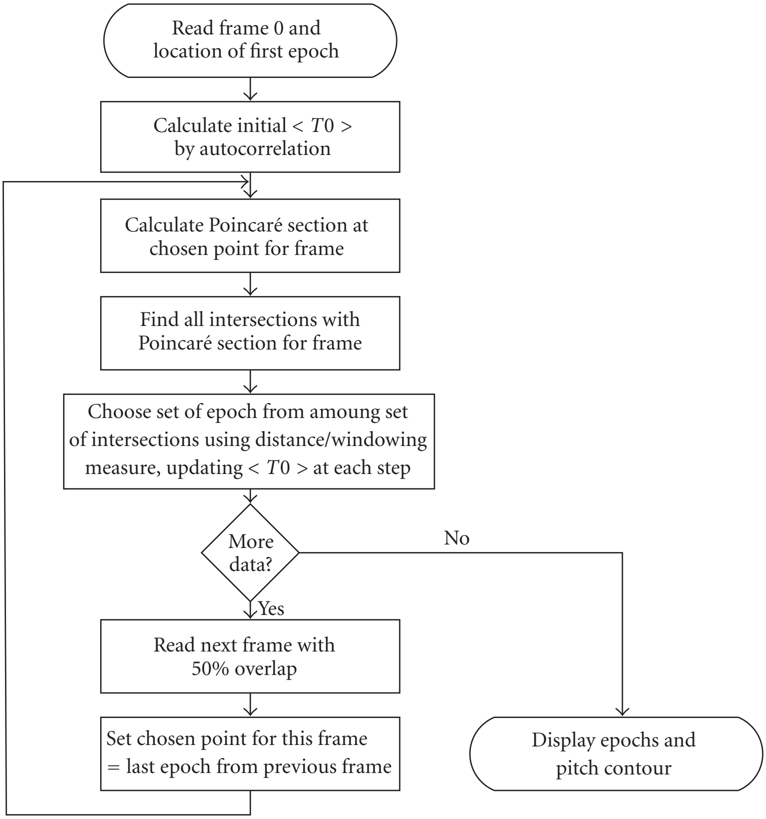

Read frame 0 and location of first epoch
Calculate initial < T0 > by autocorrelation
Calculate Poincaré section at chosen point for frame
Find all intersections with Poincaré section for frame
Choose set of epoch from amoung set of intersections using distance/windowing measure, updating < T0 > at each step
More data?
No
Yes
Read next frame with 50% overlap
Set chosen point for this frame = last epoch from previous frame
Display epochs and pitch contour

FIGURE 5.4. Schematic of the epoch marking algorithm.

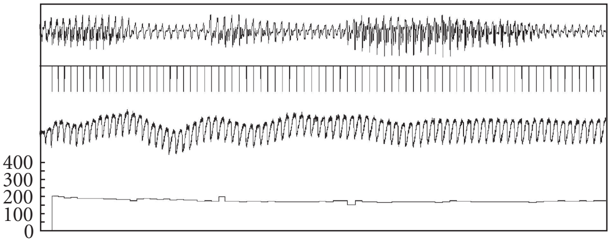


FIGURE 5.5. Results for the voiced section of "came along" from the Keele database for a female speaker. From top to bottom: the signal; the epochs as calculated by the algorithm; the laryngograph signal; and the pitch contour (Hz) resulting from the algorithm.

(1) Keele University pitch extraction database [49]. This database provides speech and laryngograph data from 15 speakers reading phonetically balanced sentences.

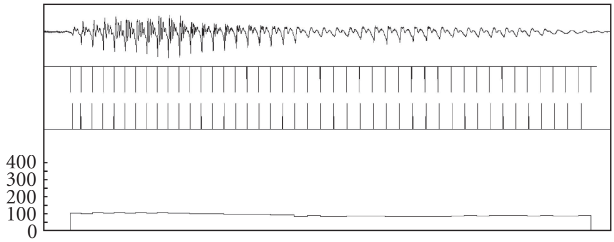


FIGURE 5.6. Results for the voiced section of "raining" from the BT Labs database for a male speaker. From top to bottom: the signal; the epochs as calculated by the algorithm; the processed laryngograph signal; and the pitch contour (Hz) resulting from the algorithm.

(2) BT Labs continuous speech. 2 phrases, spoken by 4 speakers, were processed manually to extract a data set of continuous voiced speech. Laryngograph data was also available.

The signals were up-sampled to 22.05 kHz, the BT data was originally sampled at 12 kHz, and the Keele signals at 20 kHz. All the signals had 16 bit resolution.

Figure 5.5 shows the performance of the algorithm on a voiced section taken from the phrase "a traveller came along wrapped in a warm cloak," spoken by a female speaker. There is considerable change in the signal, and hence in the attractor structure, in this example, yet the epochs are sufficiently well located when compared against the laryngograph signal.

In Figure 5.6, which is a voiced section from the phrase "see if it's raining" spoken by a male speaker, the epochs are well located for the first part of the signal, but some slight loss of synchronization can be seen in the latter part.

5.2.4. Pitch variations in LP synthesized vowels: using nonlinear methods

As was made clear in the previous section pitch modification is the key to many applications in speech. In this section we wish to consider the application of nonlinear methods to this problem. In an effort to simplify the problem, vowels generated by a linear prediction synthesizer are examined. The purpose of this exercise, which appears counter-intuitive when dealing with nonlinear synthesis, is to start with a simpler waveform, which can also be generated at any required fundamental frequency. If an analysis of this simple waveform leads to the development of a suitable algorithm for pitch modification, then this is a step towards implementing an algorithm for real speech signals.

Figure 5.7 shows the time domain waveforms and corresponding 2D projections of the 3D phase space structures for the linear prediction synthesized vowel /u/. The LP coefficients were calculated using the constant pitch /u/ vowel sound of the male speaker PB, taken from the Edinburgh 1996 database. The LP filter was then excited by a Dirac pulse train of appropriate period. This generated the stationary synthesized vowels shown, with fundamental frequencies of 70 Hz, 100 Hz, 130 Hz, and 160 Hz. These values of pitch would be typical for a male speaker.

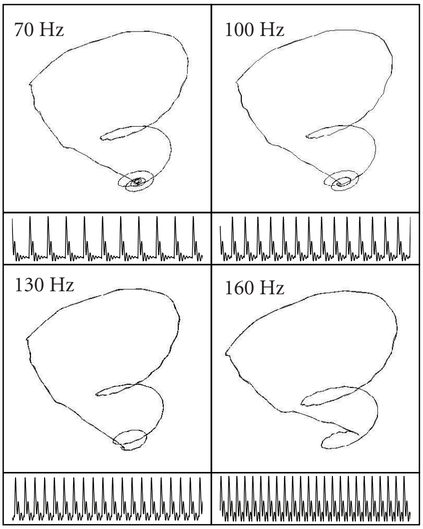


Figure 5.7. LP synthesized vowel /u/ at pitch values of 70 Hz, 100 Hz, 130 Hz, and 160 Hz.

All of these phase space reconstructions are characterized by a trajectory rising steeply out from a center point, which then returns towards this point in a series of downwards spirals. This equates to the excitation by the glottal pulse, followed by a slow decay of the waveform until the next impulse. At lower fundamental frequencies, corresponding to longer pitch periods, the number of spirals is greater, since there is a greater delay between epoch pulses. It is possible to divide each of the phase space reconstructions into two parts. The outer part consists of the outward pulse and the wide initial inward spiral, and is almost constant across all four structures. The inner part consists of the tight inner spirals close to the center point, and it is here that the variation between structures due to the pitch change can be observed. The number of inner spirals appears to be solely due to the length of the pitch period. This topological description, although applied to LP synthesized vowels, also has some similarities with the real speech signals shown previously. By a careful examination a form of outward trajectory followed by inwards spiraling can be seen.

Examining the LP phase space reconstructions from a nonlinear dynamical theory viewpoint, it would appear that there is a fixed point around which the trajectories evolve. Fixed points may take a number of forms. A more complex form is a saddle point, which has trajectories approaching the fixed point close to the inset (stable manifold) and diverging away near the outset (unstable manifold). Index 1 saddle points have an inset that is a surface, and an outset that is a curve, whereas the opposite is true of index 2 saddle points. Spiral saddle points have trajectories spiraling around the fixed point near the surface, towards or away from the saddle, for index 1 and index 2, respectively. Figure 5.8 shows an example of an index 1 spiral saddle point. Looking at the inner part of the LP phase space

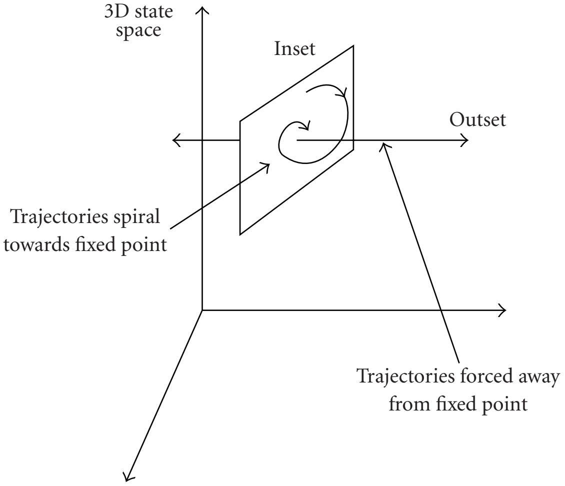

3D state space
Inset
Outset
Trajectories spiral towards fixed point
Trajectories forced away from fixed point

Figure 5.8. A spiral saddle point (index 1), with trajectories spiraling towards the fixed point near a surface and diverging away along a curve.

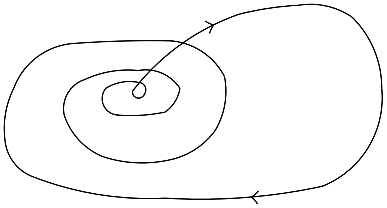


Figure 5.9. A Šilnikov-type orbit for an index 1 spiral saddle point.

reconstruction, particularly at low pitches, an index 1 spiral saddle point appears to be a good description.

A homoclinic trajectory occurs when a trajectory on the unstable manifold joins another on the stable manifold, and thus forms a single trajectory connecting the saddle to itself [46]. For spiral saddle points, this type of trajectory is also called a Šilnikov orbit, after Šilnikov's theorem [73], as shown in Figure 5.9. When the inset and outset intersect, then a so-called homoclinic intersection occurs [23]. This leads to the situation where a trajectory on the unstable manifold joins another trajectory on the stable manifold to form a single trajectory. This trajectory joins the saddle point to itself producing a homoclinic orbit.

Trajectories that come near the saddle point will approach the saddle close to the stable manifold and then be forced away from it near the unstable manifold. However, as they are pushed away by the outset, they will be recaptured by the inset and pulled back in towards the saddle point. This description captures very closely the behavior seen in all parts of the LP vowel state space reconstructions. The similarity between vowel phase space reconstructions and Šilnikov orbits has also been noted by Tokuda et al. [71], in their nonlinear analysis of the Japanese vowel /a/.

5.2.4.1. Attempted application of controlling chaos ideas

This analysis inspires a possible alternative approach to pitch modification, which operates entirely in the state space domain. It is based on concepts from the controlling chaos literature, and involves perturbing the trajectory of the speech signal in state space in order to affect a change in its orbit.

Examining once again the phase space reconstructions for different pitches of the vowel sound as produced by a linear prediction synthesizer (Figure 5.7), it appears that almost all of the information about the higher pitch sounds is contained in the lowest pitch vowel reconstruction. The effect of decreasing pitch, that is, increasing pitch period, is an increase in the number of spirals towards the center of the reconstruction, while the remainder of the phase space structure is approximately constant. Therefore it should be possible to modify the lowest pitch phase space reconstruction in some way, so that a higher pitch version can be produced. This may not be entirely realistic for real vowel sounds, but an algorithm capable of pitch modification of LP synthesized sounds would provide a stepping stone to pitch modification of real voiced speech.

5.2.4.2. Controlling chaos

In the field of nonlinear dynamics, there has been a large amount of interest in the possibility of controlling systems which exhibit chaotic behavior, so as to improve their performance. The basic principle is to locate low-period unstable periodic orbits within the attractor, which mainly comprises a large number of uncontrolled chaotic orbits. Then, using small perturbations of some control parameter, the system is moved and stabilized onto one of the low-period orbits, which is chosen so that performance is optimized. This was first proposed by Ott et al. [44], and then further refined, allowing the technique to be used with time-delay embedding, by Dressler and Nitsche [15].

5.2.4.3. Principle of pitch modification by small changes

Some of the concepts of this technique can be applied to the low-pitch vowel phase space reconstruction in order to modify the fundamental frequency. Assuming that the phase space reconstruction does have a Šilnikov-type orbit, with an index 1 saddle point at the center, then the trajectory spirals in towards the fixed point near the stable manifold (which is a plane), before being ejected out close to the unstable manifold. The process then repeats with the reinjection back towards the fixed point. The idea is to perturb the trajectory when it is close to the saddle point. Moving the trajectory closer to the plane of the stable manifold will cause it to spend longer in the region of the fixed point, thus increasing the pitch period and decreasing the pitch. Conversely, moving the trajectory away from the stable

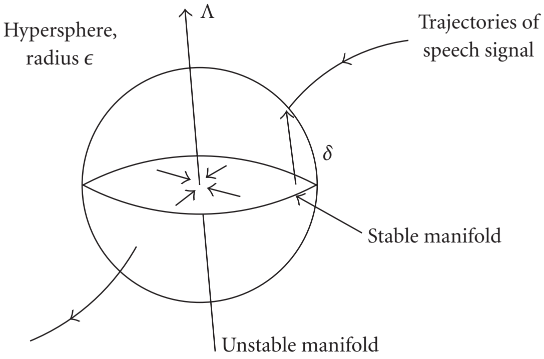


FIGURE 5.10. Principle of perturbing trajectory to modify pitch.

manifold, in the direction of the unstable manifold, will cause a faster ejection and therefore higher pitch.

An algorithm capable of perturbing the speech trajectories as described would need to perform the following operations.

(i) Embed the time series in 3D state space.

(ii) Locate the fixed point. The center of the phase space reconstruction will be close to the index 1 saddle point, where the two-dimensional stable manifold intersects with the one-dimensional unstable manifold. In practice, it will not be possible to locate the saddle point exactly, only the closest data point to it.

(iii) Find the direction of the stable and unstable manifolds. The stable manifold is expected to be a plane and the unstable manifold a curve.

(iv) Perturb the trajectory. Figure 5.10 shows the trajectory approaching the saddle point and entering a sphere of radius ϵ. At the point of entry, it is a distance δ away from the stable manifold in the direction of the unstable manifold, whose magnitude is Λ. By perturbing the trajectory towards the stable manifold (i.e., decreasing δ), the trajectory will spend longer near the fixed point, whereas moving the trajectory away from the manifold (an increase in δ) will cause a faster ejection.

(v) Calculate the relationship between the size of the perturbation and the change in pitch, so that arbitrary pitch modifications can be made.

5.2.4.4. Period modification in a Šilnikov flow

Before attempting to apply the above algorithm to the LP speech signal, modifying the period of a system which is completely specified, via a set of equations, will be examined. Consider the three coupled differential equations:

$$\begin{aligned} \dot{x} &= \alpha x - \beta y, \\ \dot{y} &= \beta x + \alpha y, \\ \dot{z} &= \gamma z. \end{aligned} \tag{5.8}$$

These define a Šilnikov flow in the region around the fixed point, although they do not model the reinjection that is characteristic of a homoclinic orbit. The system can be seen to have a fixed point at $(0, 0, 0)$, since the time derivatives go to zero at this point. Performing an eigenanalysis, the eigenvalues are found to be at

$$\begin{aligned} \lambda_1 &= \gamma, \\ \lambda_{2,3} &= \alpha \pm j\beta. \end{aligned} \tag{5.9}$$

The eigenvalue $\lambda_1 = \gamma$ has a corresponding eigenvector of $(0, 0, 1)$, and the eigenvectors of the complex conjugate eigenvalues are both in the x-y plane. If α is negative and γ is positive, then this defines an index 1 spiral saddle. Trajectories will spiral in towards the fixed point near the x-y stable manifold and then will be ejected out near the z unstable manifold.

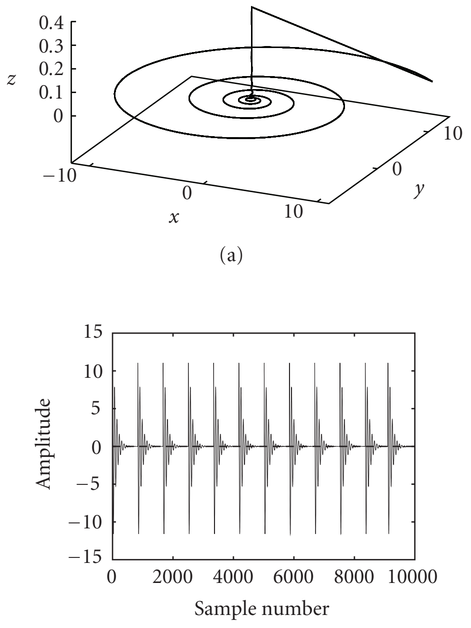
Choosing the values $\alpha = -0.1$, $\beta = 1.0$, and $\gamma = 0.08$, the equations were iterated using the fourth-order Runge-Kutta method [53], with a step size of 0.1. In order to simulate a homoclinic orbit, the trajectory was reset back to its start point after it had been ejected out along the unstable manifold a considerable distance from the fixed point (the threshold was set at $z > 0.4$). The resulting state space plot is shown in Figure 5.11(a). Plotting the variable x against time, as seen in Figure 5.11(b), results in a periodic waveform. The reinjection to complete the orbit is clearly seen. Evidently this is not very realistic, but this is not relevant as it is only the inward spiral followed by ejection that is of interest. In these terms, (5.8) with reinjection generates a realistic homoclinic orbit.

Now the principle of perturbing the trajectory is applied. To reiterate, moving the trajectory towards the stable manifold should cause an increase in the period, whereas moving away from the stable manifold (in the direction of the unstable eigenvector) will decrease the period. When the trajectory enters a sphere of radius ϵ during the Runge-Kutta iteration, it is then modified. In the experiments presented here, ϵ was set at 0.1, which is approximately 1% of the spiral radius in the x-y plane. In order to find the relationship between the period length and the modification factor, the distance from the stable manifold in the direction of the unstable manifold, δ, and the corresponding period length, n, were recorded over a range of values. Plotting n against $\log \delta$ results in a straight line, as seen in Figure 5.12. Therefore the period length n can be expressed as

$$n = a \ln \delta + b, \tag{5.10}$$

where a and b are constants, which are easily calculated from simultaneous equations. Denoting the distance from the stable manifold as the trajectory enters the sphere before modification as δ_0, then the multiplier, R, required to change the period length to n_d samples is

$$R = \frac{e^{(n_d - b)/a}}{\delta_0}. \tag{5.11}$$



(a)

(b)

Figure 5.11. Šilnikov orbit with reinjection from (5.8): (a) 3D state space and (b) x against time.

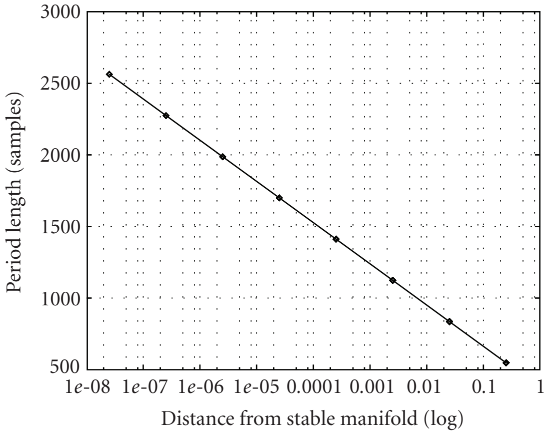


Figure 5.12. Log relationship between period length and distance from stable manifold.

Upon entering the sphere, the position vector (x, y, z) is modified to (x, y, Rz), causing a move towards or away from the stable manifold in the direction of the unstable eigenvector.

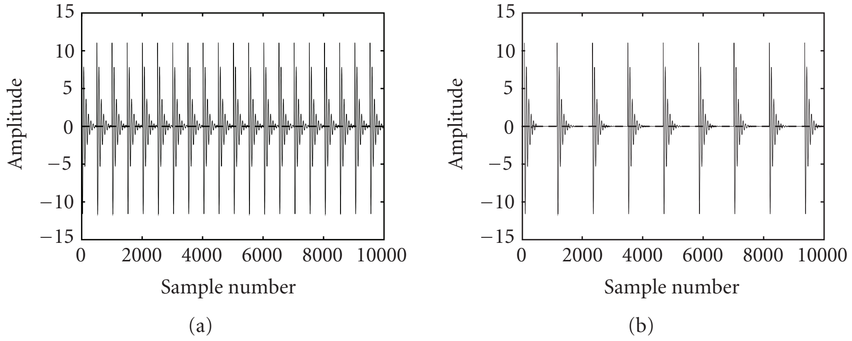


FIGURE 5.13. Period modification of the Šilnikov orbit by a factor of (a) 0.6 and (b) 1.4.

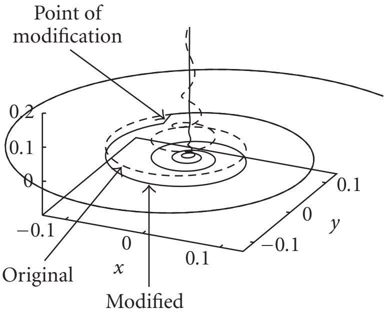


FIGURE 5.14. Zoom in on the fixed point, showing a comparison of the original and the modified trajectories.

The original orbit has a period length of 836 samples. In the following examples, the period is modified by 0.6 ($n_d = 502$ samples) and 1.4 ($n_d = 1170$ samples). a and b are calculated as -124.99 and 375.92, respectively. The values of R are then found as 14.5 and 0.069. The resulting modified waveforms are shown in Figure 5.13. Figure 5.14 shows a magnified view of the state space plot around the fixed point demonstrating the modification taking place, for the case of period modification by a factor of 1.4.

This demonstrates the validity of the trajectory perturbation approach. A Šilnikov-type orbit, which has a periodic time domain structure quite reminiscent of a vowel, has been modified, allowing both extension and shortening of the period length. In theory, it should be possible to extend the period length by any required factor by moving close to the stable manifold. The limit on period length shortening, on the other hand, is governed by the size of the sphere, since no modification occurs until the trajectory has entered it. However, it should be noted that increasing the size of the sphere beyond a small radius about the fixed point could introduce some discontinuity, and, when applied to speech, this would introduce audible artifacts.

5.2.4.5. Application to LP speech

Performing trajectory modifications when the system model is derived from a data set, rather than a set of equations, will evidently be a more complex task. The stages of the algorithm outlined in Section 5.2.4.3 and the problems found are now discussed for the LP synthesized vowel /u/.

Embedding. The LP data first has a small amount of Gaussian white noise added, to give a signal-to-noise ratio of 20 dB. This adds some variation to the signal, thus making the manifold more than a single trajectory wide. The formation of local neighborhoods can then be made by selecting near points from adjacent trajectories. The data is then embedded in three dimensions using time-delay embedding with τ set at 12 samples, equal to the first minimum of the mutual information.

Fixed point location. Because of the asymptotic nature of the fixed point, the closer the trajectory comes to it, the greater the amount of time that will be spent in the region around it. Therefore, for sampled data, the Euclidean distance between subsequent samples will decrease as the trajectory moves towards the fixed point. The data sample which has the minimum distance between adjacent samples will be the closest data point to the fixed point. This is then used as the best known position of the fixed point, $\mathbf{x}_f$, in subsequent calculations.

Direction of manifolds. An index 1 saddle point has a two-dimensional stable manifold and a one-dimensional unstable manifold. The approximate directions of these manifolds are found by an eigenanalysis of the data trajectory about the fixed point. To do this, a tangent map can be formed. Taking $\mathbf{x}_f$, the closest data point to the fixed point, as the center, a neighborhood matrix is constructed from the M points within a hypersphere around $\mathbf{x}_f$. The neighborhood matrix is then evolved forward a samples and recalculated. The tangent map then defines the linear transformation between the two local neighborhood matrices. Its eigenvalues and eigenvectors are found by SVD. Figure 5.15 shows the eigenvectors found by this method for the LP vowel /u/. The saddle point is marked and the three eigenvectors are shown. The largest eigenvector corresponds to the unstable manifold, and the two smaller eigenvectors are in the plane of the stable manifold. The parameters used in this analysis were 25 local neighbors, with an evolve length of 10 samples.

Perturbing trajectory. It is now possible to consider perturbing the trajectory, as shown in Figure 5.10, by moving the trajectory either towards or away from the stable manifold in the direction of the unstable manifold, as it enters the sphere of radius ϵ. However, an insurmountable problem is immediately encountered. Perturbing the trajectory implies moving it away from its existing course (defined by the locally linear tangent map calculated at each synthesis step) and into some

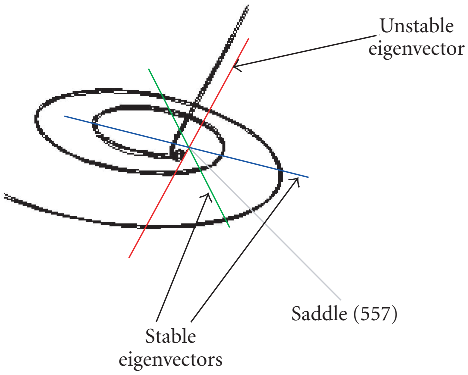


Figure 5.15. Eigenvectors around the saddle point for the LP vowel /u/.

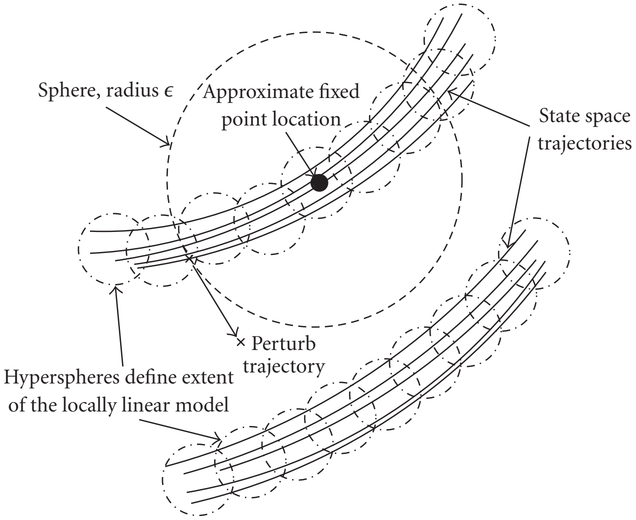

Sphere, radius ϵ
Approximate fixed point location
State space trajectories
× Perturb trajectory
Hyperspheres define extent of the locally linear model

Figure 5.16. Stylized diagram of state space around the saddle point, showing two sets of data trajectories. Perturbing trajectory implies moving into an area of state space not covered by the locally linear model.

other area of state space. This part of state space will not contain any data, and hence is not covered by the locally linear model. Therefore perturbing the trajectory moves it into a region of state space that is completely unmodeled, as shown in Figure 5.16. There is no information available to indicate how it should continue to evolve (contrary to the previous example with the set of equations that defined all points in state space), and so continuing the synthesis process after perturbing the trajectory is impossible.

This means that pitch modification by perturbing the trajectory and locally linear synthesis are not compatible, and leaves the problem of realistic pitch modification unresolved.

5.2.5. Functional approximation methods

Neural network synthesis background. Kubin and Birgmeier reported an attempt made to use a radial basis function (RBF) network approach to speech synthesis. They propose the use of a nonlinear oscillator, with no external input and global feedback in order to perform the mapping

$$x(n) = \mathcal{A}(\mathbf{x}(n-1)), \tag{5.12}$$

where $\mathbf{x}(n-1)$ is the delay vector with nonunit delays, and $\mathcal{A}$ is the nonlinear mapping function [30].

The initial approach taken [4] used a Kalman-based RBF network, which has all of the network parameters trained by the extended Kalman filter algorithm. The only parameter that must be specified is the number of centers to use. This gives good prediction results, but there are many problems with resynthesis. In particular, they report that extensive manual fine-tuning of the parameters such as dimension, embedding delay, and number and initial positions of the centers is required. Even with this tuning, synthesis of some sounds with complicated phase space reconstructions does not work [30].

In order to overcome this problem, Kubin resorted to a technique that uses all of the data points in the training data frame as centers [30]. Although this gives correct resynthesis, even allowing the resynthesis of continuous speech using a frame-adaptive approach, it is unsatisfactory due to the very large number of varying parameters, and cannot be seen as actually learning the dynamics of the speech generating system.

Following their dynamical analysis of the Japanese vowel /a/, Tokuda et al. constructed a feed-forward neural network to perform synthesis [71]. Their structure has three layers, with five neurons in the input layer, forty neurons in the hidden layer, and one in the output layer. The time delay in the input delay vector is set at $\tau = 3$ and the weights are learnt by back propagation. Using global feedback, they report successful resynthesis of the Japanese vowel /a/. The signal is noisy, but preserves natural human speech qualities. No further results in terms of speech quality or resynthesis of other vowels are given.

An alternative neural network approach was proposed by Narasimhan et al. This involves separating the voiced source from the vocal tract contribution, and then creating a nonlinear dynamical model of the source [43]. This is achieved by first inverse filtering the speech signal to obtain the linear prediction (LP) residual. Next the residue waveform is low-pass filtered at 1 kHz, then normalized to give a unit amplitude envelope. This processed signal is used as the training data in a time-delay neural network with global feedback. The NN structure reported is extremely complex, consisting of a 30 tap delay line input and two hidden layers of 15 and 10 sigmoid activation functions, with the network training performed using back propagation through time. Finally, the NN model is used in free-running synthesis mode to recreate the voiced source. This is applied to a LP filter in order to synthesize speech. They show that the NN model successfully preserves the jitter of the original excitation signal.

RBF network for synthesis. A well-known nonlinear modeling approach is the radial basis function neural network. It is generally composed of three layers, made up of an input layer of source nodes, a nonlinear hidden layer, and an output layer giving the network response. The hidden layer performs a nonlinear transformation mapping the input space to a new space, in which the problem can be better solved. The output is the result of linearly combining the hidden space, multiplying each hidden layer output by a weight whose value is determined during the training process.

The general equation of an RBF network with an input vector $\mathbf{x}$ and a single output is

$$\mathcal{F}(\mathbf{x}(n)) = \sum_{j=1}^{P} w_j \phi(\|\mathbf{x} - \mathbf{c}_j\|), \tag{5.13}$$

where there are P hidden units, each of which is weighted by w_j. The hidden units, $\phi(\|\mathbf{x} - \mathbf{c}_j\|)$, are radially symmetric functions about the point $\mathbf{c}_j$, called a center, in the hidden space, with $\|\cdot\|$ being the Euclidean vector norm [42]. The actual choice of nonlinearity does not appear to be crucial to the performance of the network. There are two distinct strategies for training an RBF network. The most common approach divides the problem into two steps. Firstly the center positions and bandwidths are fixed using an unsupervised approach, not dependent on the network output. Then the weights are trained in a supervised manner so as to minimize an error function.

Following from the work of Kubin et al., a nonlinear oscillator structure is used. The RBF network is used to approximate the underlying nonlinear dynamics of a particular stationary voiced sound, by training it to perform the prediction

$$x_{i+1} = \mathcal{F}(\mathbf{x}_i), \tag{5.14}$$

where $\mathbf{x}_i = \{x_i, x_{(i-\tau)}, \dots, x_{(i-(m-1)\tau)}\}$ is a vector of previous inputs spaced by some delay τ samples, and $\mathcal{F}$ is a nonlinear mapping function. From a nonlinear dynamical theory perspective, this can be viewed as a time-delay embedding of the speech signal into an m-dimensional state space to produce a state space reconstruction of the original d-dimensional system attractor. The embedding dimension is chosen in accordance with Takens' embedding theorem [67] and the embedding delay, τ, is chosen as the first minimum of the average mutual information function [18]. The other parameters that must be chosen are the bandwidth, the number and position of the centers, and the length of training data to be used. With these sets, the determination of the weights is linear in the parameters and is solved by minimizing a sum of squares error function, $E_S(\widehat{\mathcal{F}})$, over the N samples of training data:

$$E_s(\widehat{\mathcal{F}}) = \frac{1}{2} \sum_{i=1}^{N} (\hat{x}_i - x_i)^2, \tag{5.15}$$

where $\hat{x}_i$ is the network approximation of the actual speech signal x_i. Incorporating (5.13) into the above and differentiating with respect to the weights, then setting the derivative equal to zero gives the least-squares problem [5], which can be written in matrix form as

$$(\mathbf{\Phi}^T\mathbf{\Phi})\mathbf{w}^T = \mathbf{\Phi}^T\mathbf{x}, \tag{5.16}$$

where $\mathbf{\Phi}$ is an $N \times P$ matrix of the outputs of the centers; $\mathbf{x}$ is the target vector of length N; and $\mathbf{w}$ is the P length vector of weights. This can be solved by standard matrix inversion techniques.

Two types of center positioning strategy were considered.

(1) Data subset. Centers are picked as points from around the state space reconstruction. They are chosen pseudorandomly, so as to give an approximately uniform spacing of centers about the state space reconstruction.

(2) Hyperlattice. An alternative, data independent approach is to spread the centers uniformly over an m-dimensional hyperlattice.

Synthesis. From analysis, an initial set of parameters with which to attempt resynthesis was chosen. The parameters were set at the following values.

Bandwidth = 0.8 for hyperlattice, 0.5 for data subset; dimension = 7; number of centers = 128; hyperlattice size = 1.0; and training length = 1000.

For each vowel in the database, the weights were learnt, with the centers either on a 7D hyperlattice, or chosen as a subset of the training data. The global feedback loop was then put in place to allow free-running synthesis. The results gave varying degrees of success, from constant (sometimes zero) outputs, through periodic cycles not resembling the original speech signal and noise-like signals, to extremely large spikes at irregular intervals on otherwise correct waveforms [33].

These results implied that a large number of the mapping functions learnt by the network suffered from some form of instability. This could have been due to a lack of smoothness in the function, in which case regularization theory was the ideal solution. Regularization theory applies some prior knowledge, or constraints, to the mapping function to make a well-posed problem [21].

The selection of an appropriate value for the regularization parameter, λ, is done by the use of cross-validation [5]. After choosing all the other network parameters, these are held constant and λ is varied. For each value of λ, the MSE on an unseen validation set is calculated. The MSE curve should have a minimum indicating the best value of λ for generalization. With the regularization parameter chosen by this method, the 7D resynthesis gave correct results for all of the signals except KH /i/ and KH /u/ when using the data subset method of center selection. However, only two signals (CA /i/ and MC /i/) were correctly resynthesized by the hyperlattice method. It was found that λ needed to be increased significantly to ensure correct resynthesis for all the signals when the hyperlattice was used. Achieving stable resynthesis inevitably comes at some cost. By forcing smoothness onto the approximated function there is the risk that some of the finer detail of the state space reconstruction will be lost. Therefore, for best results, λ should be set at the smallest possible value that allows stable resynthesis. The performance of

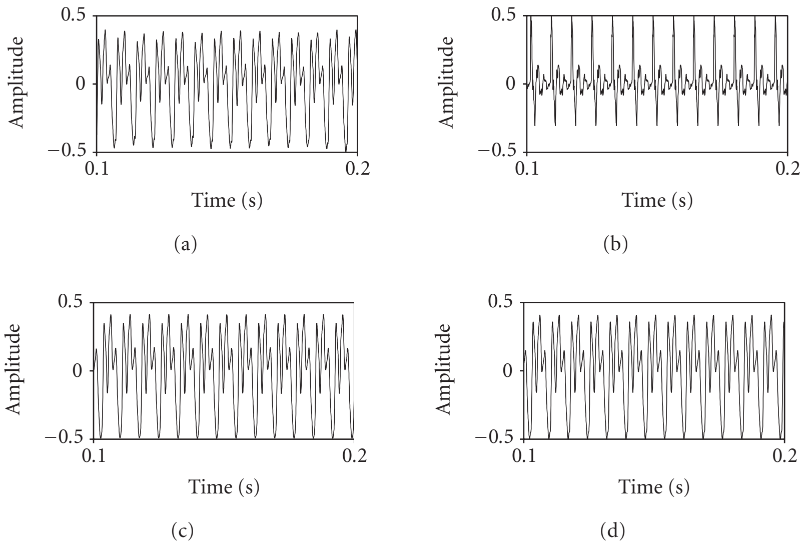
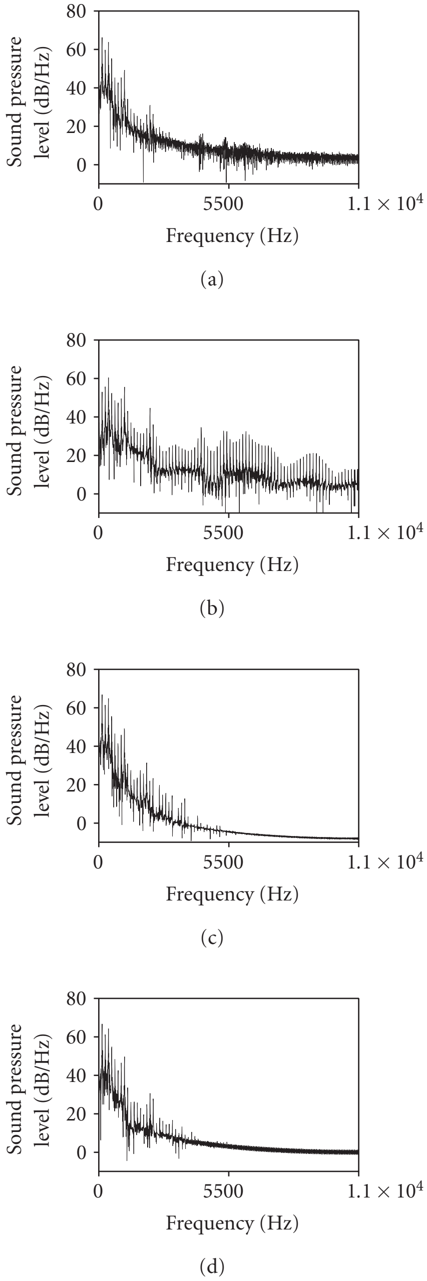


FIGURE 5.17. Time domain examples of the vowel /u/, speaker MC. (a) Original signal and (b) linear prediction synthesized signal; RBF network synthesized signal for (c) hyperlattice and (d) data subset. (Reprinted from Signal Processing, Vol.81, Lain Mann and Stephen McLaughlin, "Synthesising natural-sounding vowels using a nonlinear dynamical model," pages 1743–1756 © 2001 with permission Elsevier Science.)

the regularized RBF network as a nonlinear speech synthesizer is now measured by examining the time and frequency domains, as well as the dynamical properties. In addition to comparing the output of the nonlinear synthesizer to the original speech signal, the synthetic speech from a traditional linear prediction synthesizer is also considered. In this case, the LP filter coefficients were found from the original vowel sound (analogous to the training stage of the RBF network). The estimate ($F_s + 4$) [56] was used to set the number of filter taps to 26. Then, using the source-filter model, the LP filter was excited by a Dirac pulse train to produce the desired length LP synthesized signal. The distance between Dirac pulses was set to be equal to the average pitch period of the original signal. In this way, the three vowel sounds for each of the four speakers in the database were synthesized.

Figure 5.17 shows the time domain waveforms for the original signal, the LP synthesized signal and the two RBF synthesized signals, for the vowel /u/, speaker MC. Figure 5.18 shows the corresponding frequency domain plots of the signals, and the spectrograms are shown in Figure 5.19. In these examples, the regularization parameter λ was set at 0.01 for the hyperlattice, and 0.005 for the data subset. In the linear prediction case, the technique attempts to model the spectral features of the original, hence results in the reasonable match seen in the spectrum (although the high frequencies have been overemphasized), but the lack of resemblance in the time domain. The RBF techniques, on the other hand, resemble the original in the time domain, since it is from this that the state space reconstruction



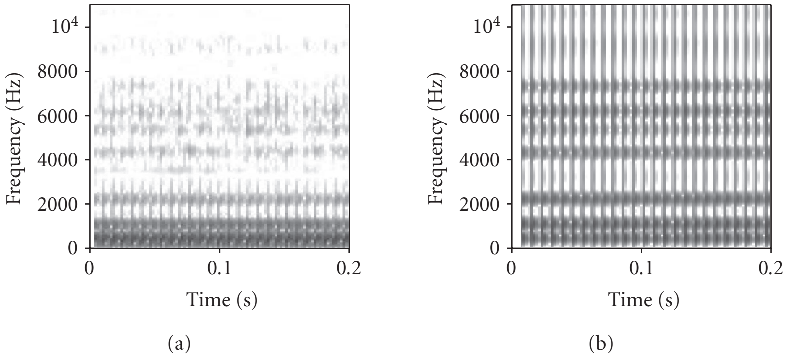
(a)

(b)

(c)

(d)

Figure 5.18. Spectrums for examples of the vowel /u/, corresponding to the signals in Figure 5.17. (a) Original signal, (b) linear prediction synthesized signal; RBF network synthesized signal for (c) hyperlattice, (d) data subset. (Reprinted from Signal Processing, Vol.81, Lain Mann and Stephen McLaughlin, "Synthesising natural-sounding vowels using a nonlinear dynamical model," pages 1743–1756 © 2001 with permission Elsevier Science.)



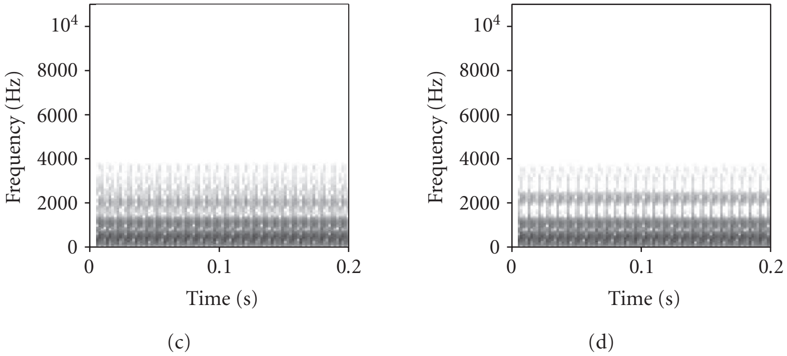


FIGURE 5.19. Wideband spectrograms for examples of the vowel /u/ corresponding to the signals in Figure 5.17. (a) Original signal, (b) linear prediction synthesized signal; RBF network synthesized signal for (c) hyperlattice, (d) data subset. (Reprinted from Signal Processing, Vol.81, Lain Mann and Stephen McLaughlin, "Synthesising natural-sounding vowels using a nonlinear dynamical model," pages 1743–1756 © 2001 with permission Elsevier Science.)

is formed, although the spectral plots show the higher frequencies have not been well modeled by this method. This is because the networks have missed some of the very fine variations of the original time domain waveform, which may be due to the regularization.

Further spectrogram examples for different vowels and speakers follow the same pattern, with the size of λ being seen to influence the quality of the signal at high frequencies.

Jitter and shimmer. Jitter and shimmer measurements were made on all of the original and RBF synthesized waveforms, using epoch detection[2] over a 500 millisecond window. Jitter is defined as the variation in length of individual pitch periods and for normal healthy speech should be between 0.1% and 1% of the average pitch period [61]. Table 5.1 shows the results of the average pitch length variation, expressed as a percentage of the average pitch period length. Results

[2] Using entropic laboratory's ESPS epoch function.

TABLE 5.1. Percentage jitter and shimmer in original and synthesized waveforms (hyperlattice and data subset) averaged over the vowels /i/, /a/, and /u/ for each speaker, and as an average over the database. (Reprinted from Signal Processing, Vol.81, Lain Mann and Stephen McLaughlin, "Synthesising natural-sounding vowels using a nonlinear dynamical model," pages 1743–1756 © 2001 with permission Elsevier Science.)

Data type	MC (male)	CA (female)	Average (female)
Hyperlattice jitter (%)	0.470	1.14	0.697
Data subset jitter (%)	0.482	0.663	0.521
Original jitter (%)	0.690	0.685	0.742
Hyperlattice shimmer (%)	1.00	1.33	0.922
Data subset shimmer (%)	0.694	7.65	2.34
Original shimmer (%)	4.21	7.06	5.17

for both center placing techniques are presented, with the jitter measurements of the original speech data. The hyperlattice synthesized waveforms contain more jitter than the data subset signals, and both values are reasonable compared to the original.

Shimmer results (the variations in energy each pitch cycle) for the original and synthesized waveforms are also displayed in Table 5.1. It can be seen that in general there is considerably less shimmer on the synthesized waveforms as compared to the original, which will detract from the quality of the synthetic speech.

Incorporating pitch into the nonlinear synthesis method. The approach adopted here is to model the vocal tract as a forced nonlinear oscillator and to embed an observed scalar time series of a vowel with pitch information into a higher dimensional space. This embedding, when carried out correctly, will reconstruct the data onto a higher dimensional surface which embodies the dynamics of the vocal tract, (see, e.g., [63, 64] for issues regarding embedding).

Previous studies, discussed above, have successfully modeled stationary (i.e., constant pitch) vowel sounds using nonlinear methods, but these have very limited use since the pitch cannot be modified to include prosody information. The new approach described here resolves this problem by including pitch information in the embedding. Specifically, a nonstationary vowel sound is extracted from a database and, using standard pitch extraction techniques, a pitch contour is calculated for the time series so that each time domain sample has an associated pitch value. In the present study measurements of rising pitch vowel sounds, where the pitch rises through the length of the time series, have been used as the basis for modeling; see, for example, Figures 5.20 and 5.21.

The time series is then embedded in an m-dimensional space, along with the pitch contour, to form an $(m + 1)$-dimensional surface. A mixed embedding delay between time samples (greater than unity) is used to capture the variable time scales present in the vowel waveform. The $(m+1)$-dimensional surface is modeled by a nearest neighbor approach, which predicts the next time series sample given

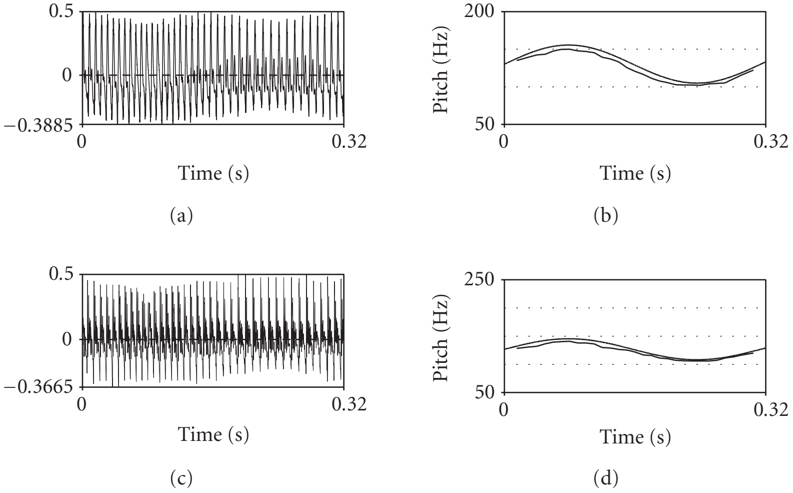


Figure 5.20. Synthesized vowel sounds together with desired and measured pitch profiles. (a) RV1, (c) RV4.

a vector of previous time samples and a pitch value (it is envisaged that more sophisticated modeling techniques will be incorporated at a later date).

Synthesis is then performed by a modification of the nonlinear oscillator approach [20], whereby the input signal is removed and the delayed synthesizer output is fed back to form the next input sample. In contrast to previous techniques, the required pitch contour is also passed into the model as an external forcing input. Our results show that this method allows the vowel sound to be generated correctly for arbitrary specified pitch contours (within the input range of pitch values), even though the training data is only made up of the rising vowel time series and its associated pitch contour. In addition, sounds of arbitrary duration can be readily synthesized by simply running the oscillator for the required length of time. Typical synthesis results are shown. It can be seen that the sinusoidal pitch contour of the synthesized sound is quite different from the rising pitch profile of the measured data; the duration of the synthesized data is also somewhat longer than that of the measured data. The small offset evident between desired and synthesized pitch contours is attributed to a minor calibration error.

The initial results presented here are encouraging. Indeed, perhaps somewhat surprisingly so since a limited measured pitch excitation data set, involving a simple rising pitch profile with a small number of data points at each specific pitch value, was used. Specifically, good synthesis results are obtained using a simple nearest neighbor embedding model with only sparse data (typically around 1000 data points embedded in a space of dimension 17, corresponding to a very low density of around only 1.5 data points per dimension).

Nonlinear function approximation models for chaotic systems. In their attempt to model and analyze nonlinear dynamics in speech signals, Kokkinos and Maragos

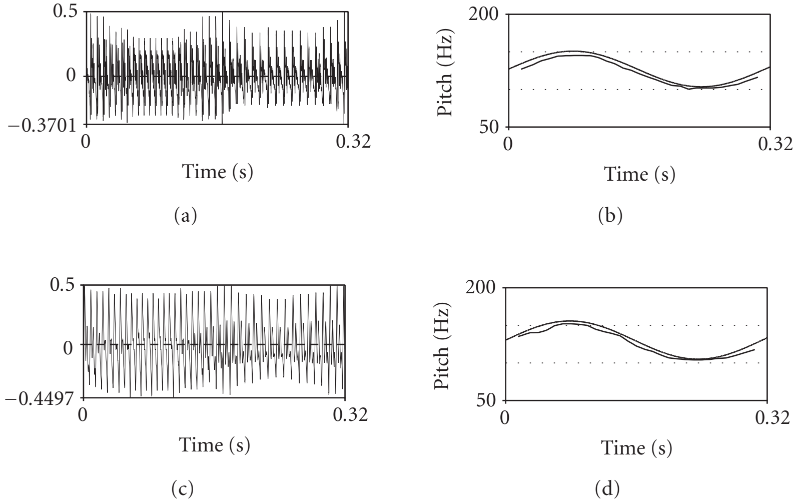


Figure 5.21. Synthesized vowel sounds together with desired and measured pitch profiles. (a) RV5, (c) RV6.

[29] have explored the applicability of nonlinear function approximation methods for the approximation of the speech production system dynamics; as in related work, the modeling is done not on the scalar speech signal, but on its reconstructed multidimensional attractor by embedding the scalar signal into a phase space. However, in contrast to the aforementioned approaches to nonlinear speech synthesis, the authors' focus has been on facilitating the application of the methods of chaotic signal analysis even when only a short-time series is available, like phonemes in natural speech utterances. This introduces an increased degree of difficulty that has been dealt with by resorting to sophisticated function approximation models that are appropriate for short data sets. A variety of nonlinear models have been explored, ranging from commonly used approximations based on global or local polynomials as well as approximations inspired from machine learning such as radial basis function networks, fuzzy logic systems and support vector machines.

Among the set of models explored, the authors opted for the use of the Takagi-Sugeno-Kang [66] model from the fuzzy logic literature, which can be seen as a special case of the probabilistic mixture of experts architecture used for function approximation. The expression used for the approximation $\widehat{\mathcal{F}}$ of the nonlinear function $\mathcal{F}$ can be written as

$$\widehat{\mathcal{F}}(X) = \frac{\sum_{i=1}^{M} \mu_i(X)\ell_i(X)}{\sum_{i=1}^{M} \mu_i(X)}, \tag{5.17}$$

where μ_i measures the degree of membership of X in the ith fuzzy set and $\ell_i(X)$ is the local model of the system dynamics for the ith fuzzy set. The term μ_i is typically expressed as a radially symmetric function $\phi(X - C_i)$ centered around point C_i

and ℓ_i are first-order polynomials in X. This expression was found experimentally to give accurate approximations of complex functions using short data records, while its probabilistic interpretation leaves open an interesting perspective for the incorporation of probabilistic information in speech synthesis.

Using this model has enabled the computation of useful features, like Lyapunov exponents, that are used to assist in the characterization of chaotic systems. Specifically, in [29] promising experimental results are reported, demonstrating the usefulness of Lyapunov exponents in the classification of speech phonemes in broad phoneme classes.

5.2.6. Nonlinear methods in speech recognition

Despite many decades of research, the current automatic speech recognition (ASR) systems still fall short from the corresponding human cognitive abilities, especially in noisy environments, because of the limitations of their acoustic processing, pattern recognition, and linguistic subsystems. Thus, there is an industrial need to develop improved robust ASR systems. Further, the complexity of the problem requires a long-term vision.

For developing the front end of ASR systems in a way consistent with the nonlinear structure of speech, one direction of nonlinear speech signal processing research has been the work of Maragos, Potamianos, and their collaborators [13, 14, 37, 48, 50]. This consists of two goals: (1) development of new and robust acoustic speech representations of the nonlinear and time-varying type (modulations and fractals) based on improved models for speech production and hearing, and (2) integration/fusion of the perceptually important among the new speech representations and the standard linear ones (cepstrum) to develop improved acoustic processing front ends for general speech recognition systems.

The motivations for the above goals include the following. (i) Adding new information to the feature set such as nonlinear and instantaneous information and good formant tracks derived from the nonlinear model can model better the aerodynamics and time evolution of speech features. (ii) Robustness to large speaker population or large vocabularies with confusable words can be achieved by using speech processing models motivated by the physics of speech production and auditory perception. (iii) Feature specialization can be achieved by investigating which of the new nonlinear features and the standard linear features correspond to the various pieces of information conveyed by the speech waveform. Such a feature tuning can lead to feature economy with corresponding reduction of computation and better acoustic modeling.

Some of the first efforts on using fractal features for ASR include the experiments in [37] on recognition of spoken letters from the E-set of the ISOLET database. Incorporating the speech fractogram as additional features to the cepstral feature vector led to a moderate decrease in the recognition error. Generalized fractal features, extracted after embedding the speech signal in a multidimensional space, have given good classification results [48] in discriminating among various sound classes, for example, fricatives, stops, vowels, and so forth, from the TIMIT database. Further, fractal-related features like the correlation dimension and the fractogram, extracted after a filtering of the nonlinear speech dynamics in the multidimensional embedding space, have yielded a notable decrease in recognition error on standard databases such as AURORA 2, especially in noisy conditions [12]. Finally, the AM-FM modulation features have given an even more significant decrease in recognition error on standard databases such as TIMIT-plus-noise and AURORA 3, as reported in [14]. It appears that these nonlinear features (of the modulation and fractal type) increase the robustness of speech recognition in noise. Ongoing work in this area deals with finding statistically optimal ways to determining the relative weights for *fusing* the nonlinear with the linear (cepstral) features.

5.3. Summary

In view of these observations, it seems likely that the data-based model of the vowel dynamics possesses an important degree of structure, perhaps reflecting physiological considerations, that requires further investigation. It is also clear that whilst encouraging there is still some way to go in overcoming the limitations of the approach. It is clear that speech is a nonlinear process and that if we are to achieve the holy grail of truly natural sounding synthetic speech than this must be accounted for. It is also clear that nonlinear synthesis techniques offer some potential to achieve this although a great deal of research work remains to be done. In the field of speech recognition, there is also strong experimental evidence that acoustic features representing various aspects of the nonlinear structure of speech can increase the robustness of recognition systems. However, more research is needed to find optimal ways for fusing the nonlinear with the linear speech features.

Acknowledgments

S. McLaughlin acknowledges the contributions of his colleagues Iain Mann, Mike Banbrook, Justin Fackrell, and Drew Lowry to this work as well as numerous useful discussions with Professor Bernie Mulgrew and Professor Dave Broomhead. P. Maragos acknowledges the contributions of his collaborators D. Dimitriadis, A. Dimakis, I. Kokkinos, V. Pitsikalis, and A. Potamianos.

Bibliography

[1] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data*, Springer, New York, NY, USA, 1996.

[2] A. Barney, C. H. Shadle, and P. O. A. L. Davies, “Fluid flow in a dynamic mechanical model of the vocal folds and tract. I. measurements and theory,” *The Journal of the Acoustical Society of America*, vol. 105, no. 1, pp. 444–455, 1999.

[3] M. C. Beutnagel, A. D. Conkie, H. J. Schroeter, Y. Stylianou, and A. K. Syrdal, “The AT&T next-gen TTS system,” in *Proceedings of Joint Meeting of ASA, EAA, and DEGA*, Berlin, Germany, March 1999.

[4] M. Birgmeier, *Kalman-trained neural networks for signal processing applications*, Ph.D. thesis, Technical University of Vienna, Vienna, Austria, 1996.

[5] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.

[6] A. C. Bovik, P. Maragos, and T. F. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3245–3265, 1993.

[7] D. M. Brookes and P. A. Naylor, "Speech production modelling with variable glottal reflection coefficient," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '88)*, vol. 1, pp. 671–674, New York, NY, USA, April 1988.

[8] D. S. Broomhead and G. P. King, "On the qualitative analysis of experimental dynamical systems," in *Nonlinear Phenonema and Chaos*, pp. 113–144, Adam Hilger, Bristol, UK, 1986.

[9] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 1805–1815, 1989.

[10] E. Moulines and R. J. Di Francesco, "Detection of the glottal closure by jumps in the statistical properties of the speech signal," *Speech Communication*, vol. 9, no. 5-6, pp. 401–418, 1990.

[11] A. G. Dimakis and P. Maragos, "Phase-modulated resonances modeled as self-similar processes with application to turbulent sounds," *IEEE Transactions on Signal Processing*, vol. 53, no. 11, pp. 4261–4272, 2005.

[12] D. Dimitriadis, N. Katsamanis, P. Maragos, G. Papandreou, and V. Pitsikalis, "Towards automatic speech recognition in adverse environments," in *Proceedings of 7th Hellenic European Conference on Research on Computer Mathematics and Its Applications (HERCMA '05)*, Athens, Greece, September 2005.

[13] D. Dimitriadis and P. Maragos, "Robust energy demodulation based on continuous models with application to speech recognition," in *Proceedings of 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp. 2853–2856, Geneva, Switzerland, September 2003.

[14] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 621–624, 2005.

[15] U. Dressler and G. Nitsche, "Controlling chaos using time delay coordinates," *Physical Review Letters*, vol. 68, no. 1, pp. 1–4, 1992.

[16] M. Edgington, A. Lowry, P. Jackson, A. P. Breen, and S. Minnis, "Overview of current text-to-speech techniques: Part II—prosody and speech generation," *BT Technical Journal*, vol. 14, no. 1, pp. 84–99, 1996.

[17] G. Fant, *Acoustic Theory of Speech Production*, Mouton, The Hague, The Netherlands, 1960.

[18] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical Review A*, vol. 33, no. 2, pp. 1134–1140, 1986.

[19] B. Gabioud, "Articulatory models in speech synthesis," in *Fundamentals of Speech Synthesis and Speech Recognition*, pp. 215–230, John Wiley & Sons, New York, NY, USA, 1994.

[20] H. Haas and G. Kubin, "A multi-band nonlinear oscillator model for speech," in *Proceedings of 32nd IEEE Asilomar Conference on Signals, Systems & Computers*, vol. 1, pp. 338–342, Pacific Grove, Calif, USA, November 1998.

[21] S. Haykin and J. C. Principe, "Making sense of a complex world [chaotic events modeling]," *IEEE Signal Processing Magazine*, vol. 15, no. 3, pp. 66–81, 1998.

[22] G. C. Hegerl and H. Hoge, "Numerical simulation of the glottal flow by a model based on the compressible Navier-Stokes equations," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '91)*, vol. 1, pp. 477–480, Toronto, Ontario, Canada, May 1991.

[23] R. C. Hilborn, *Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers*, Oxford University Press, New York, NY, USA, 1994.

[24] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal chords," *Bell System Technical Journal*, vol. 51, no. 6, pp. 1233–1268, 1972.

[25] J. F. Kaiser, “Some observations on vocal tract operation from a fluid flow point of view,” in *Vocal Fold Physiology: Biomechanics, Acoustics, and Phonatory Control*, I. R. Titze and R. C. Scherer, Eds., pp. 358–386, Denver Center for the Performing Arts, Denver, Colo, USA, 1983.

[26] D. H. Klatt, “Software for a cascade/parallel formant synthesizer,” *The Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971–995, 1980.

[27] T. Koizumi, S. Taniguchi, and S. Hiromitsu, “Glottal source—vocal tract interaction,” *The Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1541–1547, 1985.

[28] T. Koizumi, S. Taniguchi, and S. Hiromitsu, “Two-mass models of the vocal cords for natural sounding voice synthesis,” *The Journal of the Acoustical Society of America*, vol. 82, no. 4, pp. 1179–1192, 1987.

[29] I. Kokkinos and P. Maragos, “Nonlinear speech analysis using models for chaotic systems,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1098–1109, 2005.

[30] G. Kubin, “Synthesis and coding of continuous speech with the nonlinear oscillator model,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’96)*, vol. 1, pp. 267–270, Atlanta, Ga, USA, May 1996.

[31] G. Kubin, “Poincaré section techniques for speech,” in *Proceedings of IEEE Workshop on Speech Coding for Telecommunications Proceeding*, pp. 7–8, Pocono Manor, Pa, USA, September 1997.

[32] B. B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman, New York, NY, USA, 1982.

[33] I. Mann, *An investigation of nonlinear speech synthesis and pitch modification techniques*, Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland, UK, 1999.

[34] P. Maragos, “Fractal aspects of speech signals: dimension and interpolation,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’91)*, vol. 1, pp. 417–420, Toronto, Ontario, Canada, May 1991.

[35] P. Maragos, A. G. Dimakis, and I. Kokkinos, “Some advances in nonlinear speech modeling using modulations, fractals, and chaos,” in *Proceedings of 14th IEEE International Conference on Digital Signal Processing (DSP ’02)*, vol. 1, pp. 325–332, Santorini, Greece, July 2002.

[36] P. Maragos, J. F. Kaiser, and T. F. Quatieri, “Energy separation in signal modulations with application to speech analysis,” *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[37] P. Maragos and A. Potamianos, “Fractal dimensions of speech sounds: computation and application to automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1925–1932, 1999.

[38] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer, Berlin, Germany, 1976.

[39] R. McAulay and T. F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[40] R. S. McGowan, “An aeroacoustic approach to phonation,” *The Journal of the Acoustical Society of America*, vol. 83, no. 2, pp. 696–704, 1988.

[41] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[42] B. Mulgrew, “Applying radial basis functions,” *IEEE Signal Processing Magazine*, vol. 13, no. 2, pp. 50–65, 1996.

[43] K. Narasimhan, J. C. Principe, and D. G. Childers, “Nonlinear dynamic modeling of the voiced excitation for improved speech synthesis,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’99)*, vol. 1, pp. 389–392, Phoenix, Ariz, USA, March 1999.

[44] E. Ott, C. Grebogi, and J. A. Yorke, “Controlling chaos,” *Physical Review Letters*, vol. 64, no. 11, pp. 1196–1199, 1990.

[45] J. H. Page and A. P. Breen, “The Laureate text-to-speech system-architecture and applications,” *BT Technology Journal*, vol. 14, no. 1, pp. 57–67, 1996.

[46] T. S. Parker and L. O. Chua, *Practical Numerical Algorithms for Chaotic Systems*, Springer, New York, NY, USA, 1989.

[47] H.-O. Peitgen, H. Jürgens, and D. Saupe, *Chaos and Fractals: New Frontiers of Science*, Springer, New York, NY, USA, 1992.

[48] V. Pitsikalis, I. Kokkinos, and P. Maragos, "Nonlinear analysis of speech signals: generalized dimensions and lyapunov exponents," in *Proceedings of 8th European Conference on Speech Communication & Technology (EUROSPEECH '03)*, pp. 817–820, Geneva, Switzerland, September 2003.

[49] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proceedings of 4th European Conference on Speech Communication and Technology (EUROSPEECH '95)*, vol. 1, pp. 837–840, Madrid, Spain, September 1995.

[50] A. Potamianos and P. Maragos, "Time-frequency distributions for automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 196–200, 2001.

[51] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3795–3806, 1996.

[52] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an AM-FM modulation model," *Speech Communication*, vol. 28, no. 3, pp. 195–209, 1999.

[53] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, New York, NY, USA, 1992.

[54] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2001.

[55] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[56] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1978.

[57] G. Richard, M. Liu, D. Snider, et al., "Numerical simulations of fluid flow in the vocal tract," in *Proceedings of 4th European Conference on Speech Communication and Technology (EUROSPEECH '95)*, pp. 1297–1300, Madrid, Spain, September 1995.

[58] J. Schoentgen, *Dynamic models of the glottal pulse*, Elsevier, Amsterdam, The Netherlands, 1995.

[59] J. Schoentgen, "Non-linear signal representation and its application to the modelling of the glottal waveform," *Speech Communication*, vol. 9, no. 3, pp. 189–201, 1990.

[60] J. Schoentgen, "Glottal waveform synthesis with Volterra shaping functions," *Speech Communication*, vol. 11, no. 6, pp. 499–512, 1992.

[61] J. Schoentgen and R. de Guchteneere, "An algorithm for the measurement of jitter," *Speech Communication*, vol. 10, no. 5-6, pp. 533–538, 1991.

[62] C. H. Shadle, A. Barney, and P. O. A. L. Davies, "Fluid flow in a dynamic mechanical model of the vocal folds and tract. II: implications for speech production studies," *The Journal of the Acoustical Society of America*, vol. 105, no. 1, pp. 456–466, 1999.

[63] J. Stark, "Delay embeddings for forced systems. I: deterministic forcing," *Journal of Nonlinear Science*, vol. 9, no. 3, pp. 255–332, 1999.

[64] J. Stark, D. S. Broomhead, M. E. Davies, and J. P. Huke, "Takens embedding theorems for forced and stochastic systems," in *Proceedings of 2nd World Congress of Nonlinear Analysts*, Athens, Greece, July 1996.

[65] I. Steinecke and H. Herzel, "Bifurcations in an asymmetric vocal-fold model," *The Journal of the Acoustical Society of America*, vol. 97, no. 3, pp. 1874–1884, 1995.

[66] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 15, no. 1, pp. 116–132, 1985.

[67] F. Takens, "Detecting strange attractors in turbulence," in *Proceedings of Symposium on Dynamical Systems and Turbulence*, D. A. Rand and L. S. Young, Eds., vol. 898 of *Lecture Notes in Mathematics*, pp. 366–381, Coventry, UK, 1980.

[68] D. Talkin, "Voicing epoch determination with dynamic programming," *The Journal of the Acoustical Society of America*, vol. 85, no. S1, p. S149, 1989.

[69] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds., vol. 55 of *NATO Advanced Study Institute Series D*, pp. 241–261, Bonas, France, July 1989.

[70] T. J. Thomas, "A finite element model of fluid flow in the vocal tract," *Computer Speech & Language*, vol. 1, pp. 131–151, 1986.

[71] I. Tokuda, R. Tokunaga, and K. Aihara, “A simple geometrical structure underlying speech signals of the Japanese vowel /a/,” *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, vol. 6, no. 1, pp. 149–160, 1996.

[72] D. J. Tritton, *Physical Fluid Dynamics*, Oxford University Press, New York, NY, USA, 2nd edition, 1988.

[73] L. P. Šil’nikov, “A case of the existence of a denumerable set of periodic motions,” *Soviet Mathematics Doklady*, vol. 6, pp. 163–166, 1965.

Steve McLaughlin: Institute for Digital Communications, School of Engineering and Electronics, University of Edinburgh, Edinburgh EH9 3JL, Scotland, UK

Email: sml@ee.ed.ac.uk

Petros Maragos: School of Electrical and Computer Engineering, National Technical University of Athens, Athens 157 73, Greece

Email: maragos@cs.ntua.gr