Chapter 2
Nonlinearity Framework in Speech Processing

2.1 Introduction

This chapter presents a survey of nonlinear methods for speech processing. Recent developments in nonlinear science have already found their way into a wide range of engineering disciplines, including digital signal processing. It is also important and challenging to develop the nonlinear framework for speech processing because of the well known nonlinearities in the human speech production mechanism.

2.2 Nonlinear Techniques in Speech Processing

The use of nonlinear techniques in speech processing is a rapidly growing area of research. There are large variety of methods found in the literature, including linearization as in the field of adaptive filtering, introduced by Haykin [1] and various forms of oscillators and nonlinear predictors, as introduced by Kubin [2]. Nonlinear predictors are part of the more general class of nonlinear autoregressive models. Various approximations for nonlinear autoregressive models have been proposed in two main categories: parametric and nonparametric methods. In [3], Kumar et al. show how parametric methods are exemplified by polynomial approximation, locally linear models and state dependent models. Another important group of parametric methods is based on neural nets, radial basis functions approximations, as demonstrated by Birgmeier [4, 5], de Maria and Figueiras [6], and Mann and McLoughlin [7], multi-layer perceptrons as shown by Tishby [8], Wu et al. [9] and Thyssen et al. [10] and recurrent neural nets, as seen in the work of Wu et al. [9] and Hussain [11]. Nonparametric methods include various nearest neighbor methods [12] and kernel-density estimates.

Another class of nonlinear speech processing methods include models and digital signal processing algorithms proposed to analyze nonlinear phenomena of the fluid dynamics type in the speech airflow during speech production as proposed by Teager

R. S. Holambe and M. S. Deshpande, *Advances in Non-Linear Modeling for Speech Processing*, SpringerBriefs in Speech Technology,
DOI: 10.1007/978-1-4614-1505-3_2, © The Author(s) 2012

[13]. The investigation of the speech airflow nonlinearities can result in development of nonlinear signal processing systems suitable to extract related information of such phenomena. Recent work by Maragos et al., includes speech resonances modeling using AM-FM model [14]. Further, measuring the degree of turbulence in speech sounds using fractals is explained by Maragos and Potamianos in [15]. The nonlinear speech features are applied to the problem of speech recognition by Dimitriadis et al. in [16], to speech vocoders by Maragos et al. [15] and Potamianos et al. [17]. We have also applied it to the problem of speaker recognition [18, 19]. To understand various linear and nonlinear techniques used for speech processing, it is very essential to know about speech production and perception mechanisms.

2.3 Speech Production Mechanism

Speech is generated as one exhales air from the lungs while the articulators move. Thus speech sound production is a filtering process in which a speech sound source excites the vocal tract filter. The source either is periodic, causing voiced speech, or is noisy (aperiodic), causing unvoiced speech. The source of the periodicity for the former is found in the larynx, where vibrating vocal cords interrupt the airflow from the lungs, producing pulses of air. The lungs provide the airflow and pressure source for speech and the vocal cords usually modulate the airflow to create many sound variations. However, it is the vocal tract that is the most important system component in human speech production. Figure 2.1 shows the anatomy of the speech production system. The vocal tract is a tube-like passageway made up of muscles and other tissues and enables the production of different sounds. For most of the sounds, the vocal tract modifies the temporal and spectral distribution of power in the sound waves, which are initiated in the glottis.

After leaving the larynx, air from the lungs passes through the pharyngeal and oral cavities, then exits at the lips. For nasal sounds, air is allowed to enter the nasal cavity (by lowering the velum), at the boundary between the pharyngeal and oral cavities. The velum (or soft palate) is kept in a raised position for most speech sounds, blocking the nasal cavity from receiving air. During nasal sounds, as well as during normal breathing, the velum lowers to allow air through the nostrils. In the vocal tract, the tongue, the lower teeth and the lips undergo significant movements during speech production.

Figure 2.2 shows the simplified model of the vocal tract with side branches [20]. The vocal tract anatomically divides into four segments: the hypopharyngeal cavities, the mesopharynx, the oral cavity and the oral vestibule (lip tube). The hypopharyngeal part of the vocal tract consists of the supraglottic laryngeal cavity and the bilateral conical cavities of the *piriform fossa*. The mesopharynx extends from the aryepiglottic fold to the anterior palatal arch. The oral cavity is the segment from the anterior palatal arch to the incisors. The oral vestibule extends from the incisors to the lip opening [20]. In the nasal cavity, there are a number of paranasal cavities that contribute anti-resonances (zeros) to the transfer function of the vocal tract [21] and has

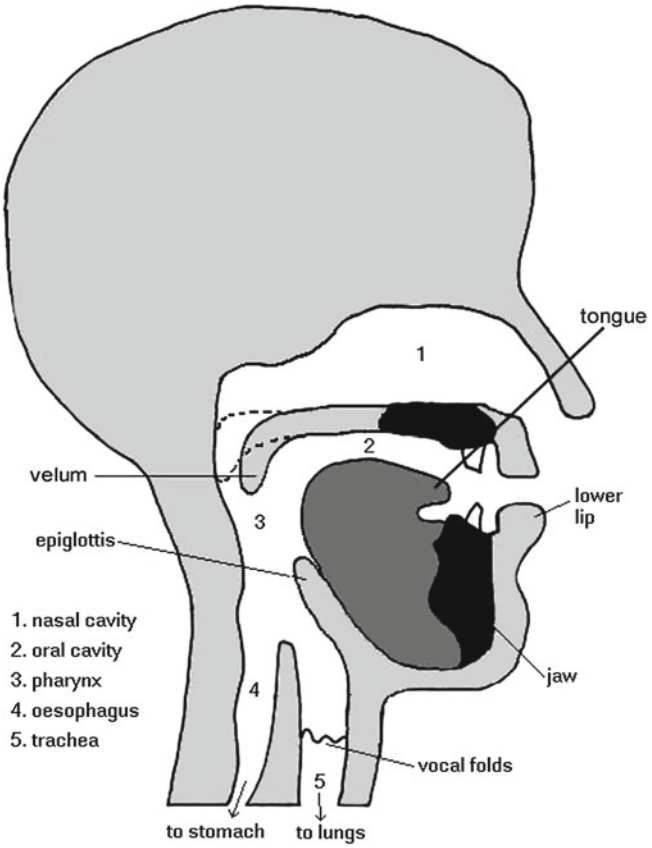

Fig. 2.1 Anatomy of human speech production system

no movable structures. Its large interior surface area significantly attenuates speech signal energy. The opening between the nasal and pharyngeal cavities controls the amount of acoustic coupling between the cavities and hence the amount of energy leaving the nostrils. Since the nasal cavity has a complicated structure and quite large individual differences, it also provides a lot of speaker-specific information. The piriform fossa is the entrance of the esophagus and is shaped like twin cone-like cavities on the left and right sides of the larynx. Because of its obscure form and function, the piriform fossa has usually been neglected in many speech production models. However, introducing the piriform fossa module into the production model causes spectral structure changes in frequency region between 4 kHz and 5 kHz, which can fit the real acoustic speech spectrum well. In addition, the piriform fossa cavities are speaker dependent and less changed during speech production. Dang and Honda suggested that, piriform fossa should be regarded as one important 'cue' for finding speaker-specific features [22]. Further they have tried to obtain such infor-

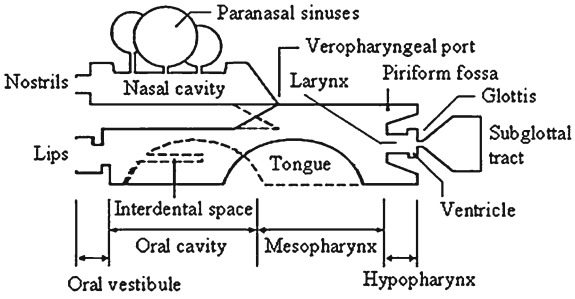

Fig. 2.2 Simplified model of vocal tract [20]

mation using MRI measurements and noted that, the hypopharyngeal resonance, i.e., the resonance of the laryngeal cavity and the antiresonance of the piriform fossa, are more stable than other formants among vowels of each speaker, while they vary to a greater extent from speaker to speaker [23, 24]. Thus the hypopharyngeal cavity also plays an important role to determine individual characteristics.

The most important aspect of speech production is the specification of different phones via the filtering actions of the vocal tract described in terms of its resonances, called formants, owing to poles in the vocal tract transfer function. The formants are often abbreviated F_i like F_1 means the formant with the lowest frequency. In voiced phones, the formants often decrease in power as a function of frequency (due to the general low pass nature of the glottal excitation); thus F_1 is usually the strongest formant. For some phones, inverse singularities of the vocal tract transfer function (zeros) exist and cause anti-resonances, where the speech power dips much more than usual between formants.

2.4 Speech Perception Mechanism

In the past, several studies have been aimed at identifying perceptual cues used by listeners, i.e., how human listener's auditory system processes speech sounds? The discipline of sound perception in general is referred to as psychoacoustics. Techniques adopted from psychoacoustics are extensively used in audio and speech processing systems for reducing the amount of perceptually irrelevant data [25].

Studies by Pickles, of the human hearing mechanism show that the processing of speech and other signals in the auditory system begins with a frequency analysis performed in the *cochlea* [26]. In the human peripheral auditory system, the input stimulus is split into several frequency bands within which two frequencies are not distinguishable. The ear averages the energies of the frequencies within each *critical band* and thus forms a compressed representation of the original stimulus. This observation has given impetus for designing perceptually motivated filter banks as front-ends for speech and speaker recognition systems.

Psychoacoustics studies have shown that human perception of the frequency content of sounds, either for pure tones or for speech signals, does not follow a linear scale. This research has led to the idea of defining subjective pitch of pure tones. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the *mel* scale. As a reference point, the pitch of 1 kHz tone, 40 db above the perceptual hearing threshold, is defined as 1000 mels. The subjective pitch in mels increases less and less rapidly as the stimulus frequency is increased linearly. The subjective pitch is essentially linear with the logarithmic frequency beyond about 1000 Hz.

Another important subjective criterion of the frequency contents of a signal is the *critical band* that refers to the bandwidth at which subjective responses, such as loudness, becomes significantly different. The loudness of a band of noise at a constant sound pressure remains constant as the noise bandwidth increases up to the width of the critical band; after that increased loudness is perceived. Similarly, a subcritical bandwidth complex sound (multitone) of constant intensity is about as loud as an equally intense pure tone of a frequency lying at the center of the band, regardless of the overall frequency separation of the multiple tones. When the separation exceeds the critical bandwidth, the complex sound is perceived as becoming louder. It shows the existence of an auditory filter in the vicinity of the tone that effectively blocks extraneous information from interfering with the detection of the tone. This vicinity is called a critical band and can be viewed as the bandwidth of each auditory filter. It is known that the width of the critical band increases with the higher frequency of the tone being masked [27]. The Bark scale is a good approximation to psychoacoustic critical band measurement.

More recently, majority of the speech and speaker recognition systems have converged to the use of feature vectors derived from a filter bank that has been designed according to some model of the auditory system. There are number of forms used for these filters, but all of them are based on a frequency scale that is roughly linear below 1 kHz and roughly logarithmic above this point. Some of the widely used frequency scales include the MEL scale [28], the BARK scale [28, 29] and the ERB (Equivalent Rectangular Bandwidth) scale [30]. In general, the peripheral auditory system can be modeled as a bank of bandpass filters, of approximately constant bandwidth at low frequencies and of a bandwidth that increases in rough proportion to frequency at higher frequencies. The popular Mel frequency cepstral coefficients (MFCCs) incorporate the MEL scale, which is represented by the following equation (since it is based on human experimental data, there are a number of approximations and models that have been used.):

$$F_{\text{Mel}} = 2595 \log_{10}\left(1 + \frac{F_{\text{Hz}}}{700}\right) \tag{2.1}$$

where F_{Hz} denotes the real frequency, and F_{Mel} denotes the perceived frequency. The Mel scale is approximately linear up to 1000 Hz and logarithmic thereafter. Another well-known mapping is the Bark-scale [28, 29]. For the Bark scale, several analytical formulae have been proposed. One of them is the one proposed by Zwicker

and Terhardt [31]:

$$F_{\text{Bark}} = 13 \tan^{-1}\left(0.76 \frac{F_{\text{Hz}}}{1000}\right) + 3.5 \tan^{-1}\left(\frac{F_{\text{Hz}}}{7500}\right)^2 \tag{2.2}$$

Another example of Bark-scale approximation is as following:

$$F_{\text{Bark}} = 6 \sinh^{-1}\left(\frac{F_{\text{Hz}}}{600}\right) \tag{2.3}$$

At the low end of the Bark scale (<1000 Hz), the bandwidths of the critical band filters are found to be about 100 Hz and in higher frequencies the bandwidths reach up to about 3000 Hz [32].

Moore and Glasberg proposed the ERB scale modifying Zwickers loudness model [30]. The ERB scale is a measure that gives an approximation to the bandwidth of filters in human hearing using rectangular bandpass filters. There are several different approximations of the ERB scale exist. The following is one of such approximations

$$\text{ERB} = 21.4 \log_{10}\left(1 + 4.37 \frac{F_{\text{Hz}}}{1000}\right) \tag{2.4}$$

Above discussion shows that, the human ear processes fundamental frequency on a logarithmic scale rather than a linear scale. Therefore, the auditory frequency analysis is most frequently modeled by a bank of bandpass filters whose bandwidths increase with increasing frequency.

2.5 Conventional Speech Synthesis Approaches

Conventional methods of speech synthesis are discussed by Mclaughlin and Maragosin in [33]. Conventionally, the 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). However, realistic articulatory synthesis is an extremely complex process and as such, instead of using it in any commercial application it is still used as 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. The resulting quality is extremely poor for voiced speech, sounding very robotic.

Concatenation methods involve joining together prerecorded units of speech which are extracted from a database. The concatenation technique provides the best quality synthesized speech. Although there is a good degree of naturalness in the synthesized output, it is still clearly distinguishable from real human speech.

McAulay and Quatieri developed a speech generation model that is based on a glottal excitation signal made up of a sum of sine waves [34]. Then, they 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.

2.6 Nonlinearity in Speech Production

Conventional theories of speech production are based on linearization of pressure and volume velocity relations and it assumed constant within a given cross section of the vocal tract, i.e., a one-dimensional planar wave assumption. We refer to this as the linear source-filter theory. While these approximations have allowed a great deal of progress to be made in understanding how speech sounds are produced and how to analyze, modify, synthesize and recognize sounds, the approximations have led to limitations. In reality, acoustic motion is not the only kind of air motion involved. The air in the vocal tract system is not static, but moves from lungs out of the mouth, carrying the sound field along with it, i.e., it contains a *nonacoustic* component. This nonacoustic phenomena, yielding a difference from the linear source-filter theory and have an impact on the *fine structure* in the speech waveform and thus how speech is processed is explained by Quatieri in [28].

The linear assumption neglects the influence of any nonacoustic motion of the fluid medium. In this model, the output acoustic pressure wave at the lips is due solely to energy from an injection of air mass at the glottis. It is known that, in this process, only a small fraction of the kinetic energy, in the flow at the glottis, is converted to acoustic energy propagated by compression and rarefaction waves [28]. The vocal tract acts as a passive acoustic filter, selectively amplifying some bands while attenuating others.

Fine structure refers to attributes in a speech waveform that can be modeled by rapid variations of parameters of traditional speech models, where rapid means on a time scale of a pitch period. For a source-filter model, the fine structure corresponds to source as well as filter. In a source-filter model, both the spacing and the amplitude of the source glottal pulses during voiced speech are considered fixed. However,

these parameters are not fixed and leads to one kind of fine structure. The variation of the fundamental period is called as *jitter* whereas period-to-period change in the pulse amplitude is called as *shimmer*. Another type of fine structure, *diplophonia*, is sometimes seen at the ends of utterances. In diplophonia, every other pitch period is both scaled down in amplitude and shifted in time. These examples of fine structure involve modifying locations and amplitudes of existing pitch pulses. Similarly, rapid variations in the filter are actually caused by interaction between the glottis (source) and the vocal tract. But the source-filter model assumes that the behavior in the glottis is not influenced by effects in the vocal tract and vice-versa. Ananthapadmanabha et al. described a model for such interactions and noted that, the effect of the glottis on the first formant is to modulate both formant frequency and the bandwidth during the open glottal phase and higher formants are less affected [35]. Such modulations are speaker-dependent to the extent that they code differences in detailed glottal behavior, resulting from physiological or other differences between speakers [36].

Even though the acoustic speech waveform and its interpretation in terms of phonetic theory is understood, models which mimic human speech production is not completely understood. Some of the difficulties in this field are related to the inadequacy of a simple source-filter model of speech production where a highly stylized source signal generator drives a slowly time-varying linear filter with negligible interaction between source and filter. Vocal fold oscillation, the turbulent sound source and the interaction phenomena are the important aspects of physical modeling of speech production and nonlinearity plays an important role in all of them. In the following, these three aspects are discussed briefly.

2.6.1 Vocal Fold Oscillation

The vocal folds, together with the aerodynamics associated to the glottis and vocal tract, constitute a self-excited biomechanical oscillator that acts as the sound source during voice production. Under certain instability conditions for the biomechanical parameters such as air pressure, vocal fold tension and glottal area, the air flow through the glottis causes the oscillation, which in turn produces the air pressure wave perceived as voice [37, 38]. Thus, it is a self-excited flow-induced oscillation, which is the same phenomenon that produces the oscillation of buildings by action of the wind, the vibration of airplane wings during flight and the generation of sound in wind musical instruments [39]. This oscillator has a relatively complex dynamical structure, as consequence of nonlinear viscoelastic characteristics of its tissues, collisions between the opposite vocal folds and nonlinear interaction between the airflow and the glottal area. Using mathematical models of that structure, past works by Lucero, [40] and Herzel et al. [41], have shown the existence of several nonlinear phenomena, such as multiple equilibrium positions and limit cycles and several types of bifurcations and chaotic behavior.

Many of the acoustic and perceptual features of an individual's voice are believed to be due to specific characteristics of the quasi-periodic excitation signal provided

by the vocal folds. These, in turn, depends on the morphology of the voice organ, the larynx. The anatomy of the larynx is quite complicated and its descriptions may be found in the literature [42]. From an engineering point of view, the larynx is the structure that houses the vocal folds whose vibration provides the periodic excitation. The space between the vocal folds, called the glottis, varies with the motion of the vocal folds and thus modulates the flow of air through them. We now know that the larynx is a self-oscillating acousto-mechanical oscillator. This oscillator is controlled by several groups of tiny muscles housed in the larynx. Some of these muscles control the rest position of the folds, others control their tension and still others control their shape. During breathing and production of fricatives, for example, the folds are pulled apart to allow free flow of air. To produce voiced speech, the vocal folds are brought close together. When brought close enough together, they go into a spontaneous periodic oscillation. These oscillations are driven by Bernoulli pressure (the same mechanism that keeps airplanes aloft) created by the airflow through the glottis. If the opening of the glottis is small enough, the Bernoulli pressure due to the rapid flow of air is large enough to pull the folds toward each other, eventually closing the glottis. This, of course, stops the flow and the laryngeal muscles pull the folds apart. This sequence repeats itself until the folds are pulled far enough away or if the lung pressure becomes too low. Besides the laryngeal muscles, the lung pressure and the acoustic load of the vocal tract also affect the oscillation of the vocal folds. These oscillations are driven from an almost stationary lung pressure. Linear time-invariant systems are unable to produce such oscillations. If we exclude a hypothetical time-varying nervous control input, it results into the conclusion that, the oscillation process is nonlinear. This nonlinearity is routinely included even in simple methods of vocal fold behavior as explained by Flanagan in [43], where it is attributed to nonlinear feedback via the Bernoulli force.

The qualitative changes in the type of steady-state motion of a nonlinear dynamical model such as the vocal folds are referred to as *bifurcations*. They show up as discontinuities when a system parameter is moved across some threshold. For instance, an equilibrium state (e.g. open glottis in unvoiced speech) may bifurcate to a periodic limit-cycle motion (e.g. after a transition to voiced speech). It has become popular to characterize vocal fold models in terms of their bifurcation diagrams [44, 45]. The transition between the model and falsetto registers is also considered a bifurcation [46]. Chaos refers to the steady-state motion of nonlinear dynamical systems characterized by high sensitivity to initial conditions. These motions often appear irregular and initial perturbations diverge exponentially.

Because of nonlinear behavior of vocal fold oscillations, it can be seen that the spectral envelop of the glottal pulse changes with its pitch frequency and the spectral content changes with its amplitude.

2.6.2 The Turbulent Sound Source

Turbulence is the source of noise-like sound in speech. In the linear speech model this has been dealt with by having a white noise source exciting the vocal tract filter. Turbulence is the dominant source for frication, aspiration and whisper, and a partial source in breathy voice and creaky voice. The common picture is that turbulence is a nonlinear phenomenon with strong interaction between the airflow and the acoustic sound field occurring at constrictions and obstacles. Turbulence is one of the prominent examples where both theoretical explanation and experimental evidence for chaos are available [47].

Turbulent airflow shows highly irregular fluctuations of particle velocity and pressure. These fluctuations are audible as broadband noise. Turbulent excitation occurs mainly at two locations in the vocal tract: near the glottis and at constriction(s) between the glottis and the lips. Turbulent excitation at a constriction downstream of the glottis produces fricative sounds or voiced fricatives depending on whether or not voicing is simultaneously present. Measurements and models for turbulent excitation are even more difficult to establish than for the periodic excitation produced by the glottis because, usually, no vibrating surfaces are involved. Because of the lack of a comprehensive model, much confusion exists over the proper sub-classification of fricatives. The simplest model for turbulent excitation is a *nozzle* (narrow orifice) releasing air into free space.

Experimental work has shown that half (or more) of the noise power generated by a jet of air originates within the so-called mixing region that starts at the nozzle outlet and extends as far as a distance four times the diameter of the orifice. The noise source is therefore distributed. Distributed sources of turbulence can be modeled by expanding them in terms of monopoles (i.e., pulsating spheres), dipoles (two pulsating spheres in opposite phase), quadrupoles (two dipoles in opposite phase) and higher-order representations. A much stronger noise source is created when a jet of air hits an obstacle. Depending on the angle between the surface of the obstacle and the direction of flow, the surface roughness and the obstacle geometry, the noise generated can be up to 20 dB higher than that generated by the same jet in free space. Because of the spatially concentrated source, modeling obstacle noise is easier than modeling the noise in a free jet. Experiments reveal that obstacle noise can be approximated by a dipole source located at the obstacle. The above theoretical findings qualitatively explain the observed phenomenon that the fricatives “th” and “f” (and the corresponding voiced “dh” and “v”) are weak compared to the fricatives “s” and “sh”. The teeth (upper for “s” and lower for “sh”) provide the obstacle on which the jet impinges to produce the higher noise levels. A fricative of intermediate strength results from a distributed obstacle when the jet is forced along the roof of the mouth.

Modern theories that attempts to explain turbulence predict the existence of eddies (vortices with a characteristic size λ) at multiple scales [48]. 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. This multiscale structure of turbulence can in some cases be quantified by *fractals*.

2.6.3 Interaction Phenomenon

The linear, one-dimensional acoustic model is too tightly constrained to accurately model many characteristics of vocal tract. The widely used linear predictive cepstral coefficient (LPCC) and Mel frequency cepstral coefficient (MFCC) features are based on this linear speech production model and assume that the airflow propagates in the voca1 tract as a linear plane wave. There is an increasing collection of evidence suggesting that nonaciustic fluid motion can significantly influence the sound field. For example, measurements by Teager reveal the presence of separated flow within the vocal tract [49]. Separated flow occurs when a region of fast moving fluid-a jet-detaches from regions of relatively stagnant fluid. When this occurs, viscous forces (neglected by linear models) create a tendency for the fluid to 'roll up' into rotational fluid structures commonly referred to as *vortices* as shown in Fig. 2.3b. Teager suggested that the presence of traveling vortices, 'smoke rings' could result in additional acoustic sources throughout the vocal tract. This contribution of non-linear excitation sources is something neglected by source-filter theory [50]. Figure 2.3a and b show the linear and nonlinear models of sound propagation along the vocal tract, respectively [28, 51].

Motivated by the measurements of Teager, Kaiser hypothesized that the interaction of the jet flow and the vortices with the vocal tract cavity is responsible for much of the speech fine structure, particularly at high formant frequencies. Then he proposed the need for time frequency analysis methods with greater resolution than short-time Fourier transform (STFT) for measuring fine structure within a glottal cycle. He further argues that the instantaneous formant frequencies may be more important than the absolute spectral shape.

2.7 Common Signals of Interest

Signals that we often use in this book are defined in this section. In Chaps. 4 and 5, we will see how some of the parameters of these signals can be estimated.

2.7.1 AM Signals

An amplitude modulated (AM) signal is the combination of two signals, where one signal is the *carrier*, which is a single frequency sinusoidal signal and the other is the information we want to transmit, i.e., the *baseband* signal. The amplitude modulated

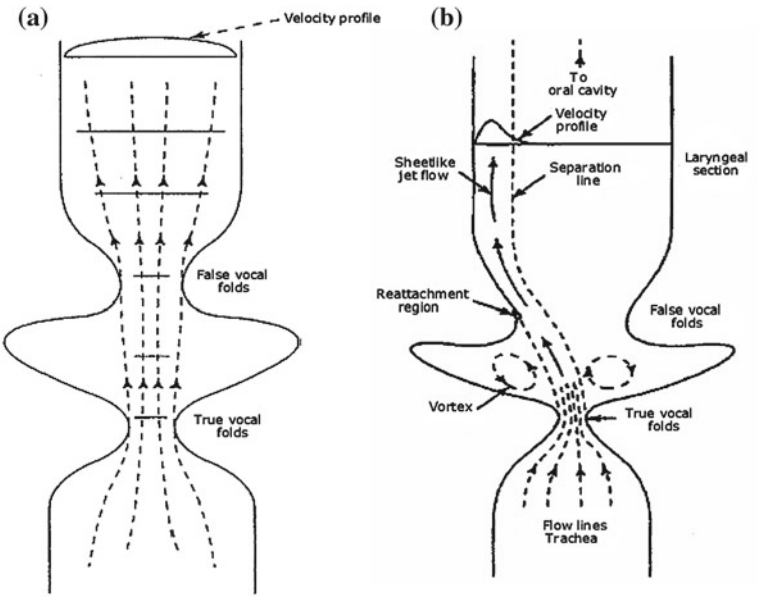

Fig. 2.3 **a** Linear and **b** nonlinear model of sound propagation along the vocal tract [28, 51]

signal can be modeled as:

$$a(t) = A[1 + km(t)] \tag{2.5}$$

$$s_{AM}(t) = a(t)cos(\Omega_c t) \tag{2.6}$$

where A is the signal amplitude, Ω_c is the carrier frequency (in radians/second), $m(t)$ is the baseband signal and k is the modulation index.

2.7.2 FM Signals

Just like the AM signals, an FM signal is the combination of two signals, where one is the single frequency sinusoidal signal, the carrier and the other is the baseband signal, however, in FM signals the baseband signal is used to change the frequency of the carrier signal.

A frequency modulated signal (FM signal) can be modeled as,

$$\phi(t) = \Omega_c t + \Omega_m \int_0^t q(\tau)d\tau + \theta \tag{2.7}$$

$$s_{FM}(t) = A\,cos(\phi(t)) \tag{2.8}$$

where A is the signal amplitude, Ω_c is the carrier frequency, Ω_m is the maximum frequency deviation with $\Omega_m \in [0, \Omega_c]$, $q(t)$ is the baseband signal with $|q(t)| \leq 1$, and θ is the constant phase offset. The instantaneous frequency is defined as the derivative of $\phi(t)$:

$$\Omega_i(t) = \frac{d\phi(t)}{dt} = \Omega_c + \Omega_m q(t) \tag{2.9}$$

2.7.3 AM-FM Signals

The AM-FM signal is the combination of both the AM and FM signals discussed above and it can modeled as,

$$\begin{aligned} s_{AM-FM}(t) &= a(t)cos[\phi(t)] \\ &= a(t)cos\left(\Omega_c t + \Omega_m \int_0^t q(\tau)d\tau + \theta\right) \end{aligned} \tag{2.10}$$

This signal can model the time-varying amplitude and frequency patterns in speech resonances. $s_{AM-FM}(t)$ is a cosine of carrier frequency Ω_c with a time-varying amplitude signal $a(t)$ and a time varying instantaneous frequency signal $\Omega_i(t)$.

2.7.4 Discrete Versions

We can get discrete versions of the AM, FM, and AM-FM signals above by sampling them. We can derive new expressions for these if we substitute t by nT and Ω by ω/T, where ω is the digital frequency (in radians/sample), and T is the sampling period. Finally, the integrations are replaced by sums.

2.8 Summary

This chapter described the speech production and perception mechanisms considering the nonlinearities present in them. Some important aspects of physical modeling of speech production system like vocal fold oscillations, the turbulent sound source, aerodynamics observations regarding nonlinear interactions between the air flow and the acoustic field are discussed in this chapter.

References

1. Haykin S (2001) Adaptive filter theory. Prentice Hall, Upper Saddle River
2. Kubin G (1995) Nonlinear processing of speech. In: Kleijn WB, Paliwal KK (eds) Speech coding and synthesis. Elsevier Science, Amsterdam
3. Kumar A, Mullick SK (1996) Nonlinear dynamical analysis of speech. J Acoust Soc Amer 100:615–629
4. Birgmeier M (1995) A fully Kalman-trained radial basis function network for nonlinear speech modeling. In: Proceedings of IEEE international conference on neural networks (ICNN'95), Perth
5. Birgmeier M (1996) Nonlinear prediction of speech signals using radial basis function networks. In: EUSIPCO'96, vol 1, pp 459–462
6. de Maria FD, Figueiras AR (1995) Nonlinear prediction for speech coding using radial basis functions. In: Proceedings of IEEE international conference on acoustics, speech, and, signal processing (ICASSP'95), pp 788–791
7. Mann I, McLaughlin S (1999) Stable speech synthesis using recurrent radial basis functions. In: Proceedings of European conference on speech communication and technology, vol 5, pp 2315–2318
8. Tishby N (1990) A dynamical systems approach to speech processing. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP'90)
9. Wu L, Niranjan M, Fallside F (1994) Fully vector quantized neural network-based code-excited nonlinear predictive speech coding. IEEE Trans Speech Audio Process 2(4)
10. Thyssen J, Nielsen H, Hansen SD (1994) Non-linear short term prediction in speech coding. In: Proceedings of IEEE international conference on acoustics, speech, and signal processing (ICASSP'94), Australia, pp I-185–I-188
11. Hussain A (1996) Novel artificial neural-network architectures and algorithms for non-linear dynamical system modelling and digital communications applications. Ph.D. thesis, University of Strathclyde, Glasgow
12. Farmer JD, Sidorowich JJ (1988) Exploiting chaos to predict the future and reduce noise. In: Lee YC (ed) Evolution, learning, and cognition. World Scientific, Singapore, pp 277–330
13. Teager HM, Teager SM (1989) Evidence for nonlinear sound production mechanisms in the vocal tract. In: Hardcastle W, Marchal A (eds) Speech production and speech modeling, vol 55. NATO Advanced Study Institute Series D, Bonas, France
14. Maragos P, Kaiser JF, Quatieri TF (1993) Energy separation in signal modulations with application to speech analysis. IEEE Trans Signal Process 41(10):3024–3051
15. Maragos P, Potamianos A (1999) Fractal dimensions of speech sounds: computation and application to automatic speech recogntion. J Acoust Soc Amer 105:1925–1999
16. Dimitriadis D, Maragos P, Potamianos A (2002) Modulation features for speech recognition. In: Proceedings of IEEE international conference on acoustics, speech, and, signal processing (ICASSP'02), pp I-377–I-380
17. Potamianos A, Maragos P (1999) Speech processing applications using an am-fm modulation model. Speech Commun 28:195–209
18. Deshpande MS, Holambe RS (2009) Speaker identification based on robust am-fm features. In: Proceedings of second IEEE international conference on emerging trends in engineering and technology (ICETET'09), Nagpur, pp 880–884
19. Deshpande MS, Holambe RS (2009) Robust q features for speaker identification. In: Proceedings of IEEE international conference on advances in recent technologies in communication and computing (ARTCom'09), Kottayam, Kerala, pp 209–213
20. Honda K (2008) Physiological processes of speech production. Springer, Berlin
21. Kitamura T, Honda K, Takemoto H (2005) Individual variation of the hypopharyngeal cavities and its acoustic effects. Acoust Sci Tech 26:16–26
22. Dang J, Honda K (1997) Acoustic characteristics of the piriform fossa in models and humans. J Acoust Soc Am 101(1):456–465

23. Kitamura T, Takemoto H, Adachi S, Mokhtari P, Honda K (2006) Cyclicity of laryngeal cavity resonance due to vocal fold vibration. J Acoust Soc Am 120(6):2239–2249
24. Dang J, Honda K (1996) An improved vocal tract model of vowel production implementing piriform fossa resonance and transvelar nasal coupling. In: Proceedings of international conference on spoken language processing (ICSLP'96), pp 965–968
25. Kinnunen T (2003) Spectral features for automatic text-independent speaker recognition. Ph.D. thesis, Finland
26. Pickles J (1982) An introduction to the physiology of hearing. Academic Press, London
27. Fletcher H (1940) Auditory patterns. Rev Mod Phys 12:47–65
28. Quatieri TF (2004) Discrete-time speech signal processing. Principles and practice. Pearson Education, London
29. Gold B, Morgan N (2002) Speech and audio signal processing. Wiley, New York
30. Moore BCJ, Glasberg BR (1996) A revision of Zwickers loudness model. Acustica–Acta Acustica 82:335–345
31. Zwicker E, Terhardt E (1980) Analytical expressions for critical band rate and critical bandwidth as a function of frequency. J Acoust Soc Am 68:1523–1525
32. Green DM (1976) An introduction to hearing. Wiley, New York
33. Mclaughlin S, Maragos P Nonlinear methods for speech analysis and synthesis. In: Marshall S, Sicuranza GL (eds) Advances in nonlinear signal and image processing. Hindawi Publishing Corporation, New York
34. McAulay RJ, Quatieri TF (1986) Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans Acoustic Speech Signal Process 34:744–754
35. Ananthapadmanabha TV, Fant D (1982) Calculation of true glottal flow and its components. Speech Commun 1:167–184
36. Jankowski CR (1996) Signal processing using the Teager energy operator and other nonlinear operators. Ph.D. thesis. MIT, Cambridge
37. Titze I (1988) The physics of small-amplitude oscillation of the vocal folds. J Acoust Soc Amer 83:1536–1552
38. Titze I (1994) Principles of voice production. Prentice-Hall, Englewood Cliffs
39. Thompson J, Stewart H (1996) Nonlinear dynamics and chaos. Wiley, New York
40. Lucero J (1999) A theoretical study of the hysteresis phenomenon at vocal fold oscillation onset-offset. J Acoust Soc Amer 105:423–431
41. Herzel H, Knudsen C (1995) Bifurcation in a vocal fold model. Nonlin Dyn 7:53–64
42. Zemlin WR (1968) Speech and hearing science, anatomy, and physiology. Prentice-Hall, Englewood Cliffs
43. Flanagan JL (1872) Speech analysis synthesis and perception, 2nd edn. Springer, New York
44. Lucero J (1993) Dynamics of the two-mass model of the vocal folds: equilibria, bifurcations and oscillation region. J Acoust Soc Amer 94:3104–3111
45. Awrejcewicz J (1991) Bifurcations and chaos in coupled oscillators. World Scientific, Singapore
46. McGowan RS (1993) The quasi-steady approximation in speech production. J Acoust Soc Amer 84:3011–3013
47. Manneville P (1990) Dissipative structures and weak turbulence. Academic Press, Boston
48. Tritton DJ (1988) Physical fluid dynamics, 2nd edn. Oxford University Press, New York
49. Teager HM (1980) Some observations on oral air flow during phonation. IEEE Trans Speech Audio Process 28(5):599–601
50. Hansen JHL, Ceballos GC, Kaiser JF (1998) A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment. IEEE Trans Biomed Eng 45(3):300–313
51. Zhou G, Hansen JHL, Kaiser JF (2001) Non-linear feature based classification of speech under stress. IEEE Trans Speech Audio Process 9(3):201–216

http://www.springer.com/978-1-4614-1504-6

Advances in Non-Linear Modeling for Speech Processing

Holambe, R.S.; Deshpande, M.S.

2012, XIII, 102 p. 32 illus., Softcover
ISBN: 978-1-4614-1504-6