CHAPTER 2- SPEECH PRODUCTION AND EVIDENCES OF NONLINEAR BEHAVIOUR

2.1 INTRODUCTION

Speech production is a complex feedback process in which also hearing, perception and information processing in the nervous system and the brain is involved. In this chapter the human organs which are contributing to the speech production process are explained in brief. For the different categories of speech sounds the mechanism of production is explained with respect to sound excitation and the different articulators. On the basis of proper tube models, the physics and acoustics of speech sound generation, propagation within the speech production system and radiation is discussed.

Voice production can be thought of as the activation of an entire system of coupled oscillators. The intent to vocalize activates motor commands that are responsible for the neural inputs to an array of biomechanical, neural, and acoustic oscillators (large box in Figure 2.1).

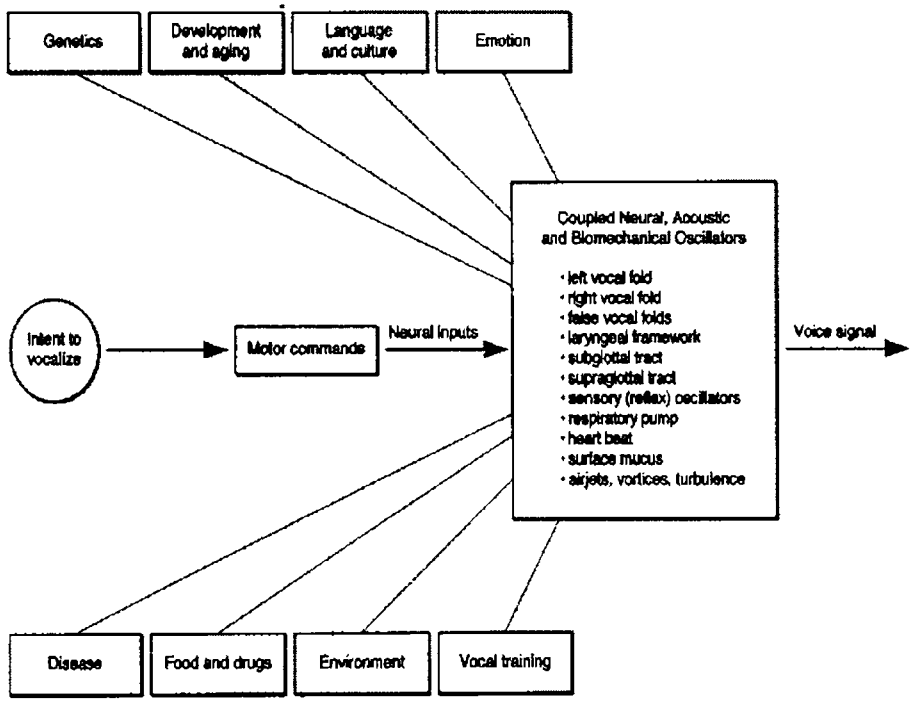

Figure 2.1. A list of biological oscillators involved in voice production and factors that may influence them.

The vocal folds are the primary oscillating system that produce the carrier signal (the glottal airflow). All other oscillators can then be thought of as modulators of the carrier signal. Some of the modulations are nearly sinusoidal (respiratory, heart beat) but many are high dimensional (action potentials of muscles, air vortices, mucus in motion). Yet others are passive oscillators (tracheal resonator, supraglottal vocal tract, various sinuses) that can influence the primary oscillating system. We can assume that the system of coupled oscillators contains and releases information about the human body; in particular, about its genetics, development, age, disease, language, culture, food and drug intake, and response to the environment (Figure 2.1).

Voice perturbation analysis has the goal of extracting some of this information from the voice signal. In all cases, the procedure is extremely difficult and usually requires considerable a priori knowledge about the modulations to be extracted. Therein lies the primary problem of voice perturbation analysis in its present state. Many studies are needed to isolate the individual contributions of each oscillator. Some of these studies are underway (J. van den Berg [1]).

Before discussing the acoustic theory and modern techniques in nonlinear dynamics for speech analysis, it is important to consider the various types of sounds that make up human speech. Speech can be broken down in to small segments called phonemes, each of which is unambiguously distinguishable and can be represented by any of a number of different phonetic alphabets. The figure 2.2 shows a schematic diagram of the vocal system .

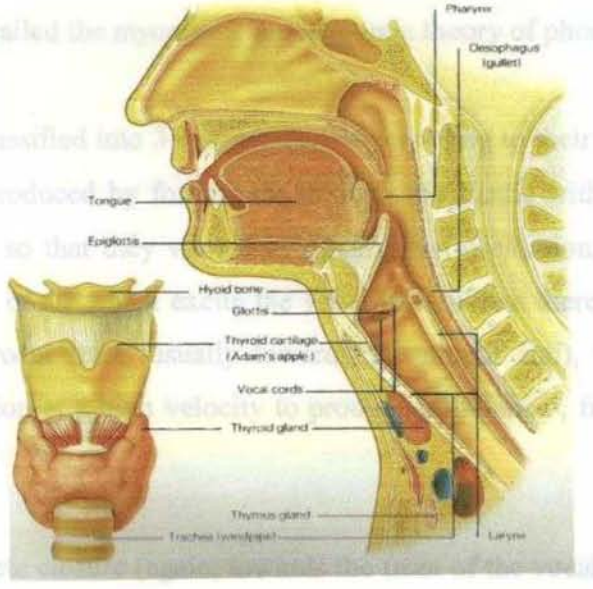

Figure 2.2. Schematic diagram of the vocal system shows components in the airway system in the head, neck, and chest (Titze, 1994a)[2]: The lungs produce pressure that drives the subglottal airstream. The airstream is fed to the larynx via bronchi and trachea. The primary function of the larynx is to protect the airway system from foreign material (such as food) which passes into the esophagus during swallowing. The vocal tract, comprised of pharynx and oral and nasal cavities, filters the primary sound signal generated by airstream-driven oscillations of the vocal folds in the larynx.

The basic mechanism for producing any sound is to expel air from the lungs, using muscular action. The actual mechanism by which we create a phoneme can be split into two main categories, voiced and unvoiced that can be further split in to vowels, fricatives or plosives.

Voiced speech or phonation is produced by oscillating the fleshy membranes inside the larynx which are known as the vocal folds. The oscillation is set up by forcing the vocal folds closed which causes pressure to build up below the folds, gradually forcing them to open again allowing the air to flow from the sub glottal region in to the mouth. This rapid air flow creates a Bernoulli force which coupled with the muscular action of the vocal muscles produces the sound.

Thus an oscillation is set up with the fundamental frequency being a function of the vocal fold tension which is controlled by the vocalis muscles. This theory of how the oscillations occur is called the myoelastic/aerodynamic theory of phonation [3].

Speech sounds are classified into 3 distinct classes according to their mode of excitation. Voiced sounds are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing quasi-periodic pulses of air which excite the vocal tract. When there is a constriction at some point in the vocal tract (usually towards the mouth end), thereby forcing air through the constriction at a high velocity to produce turbulence , fricative or unvoiced sounds are generated.

When there is complete closure (again, towards the front of the vocal tract), building up pressure behind the closure , and abruptly releasing it plosive sounds is produced.

2.2 PRODUCTION OF SPEECH

As sound generated as mentioned above, propagates down the non uniform cross-sectional tubes of the vocal tract and nasal tract shown as in figure 2.3.

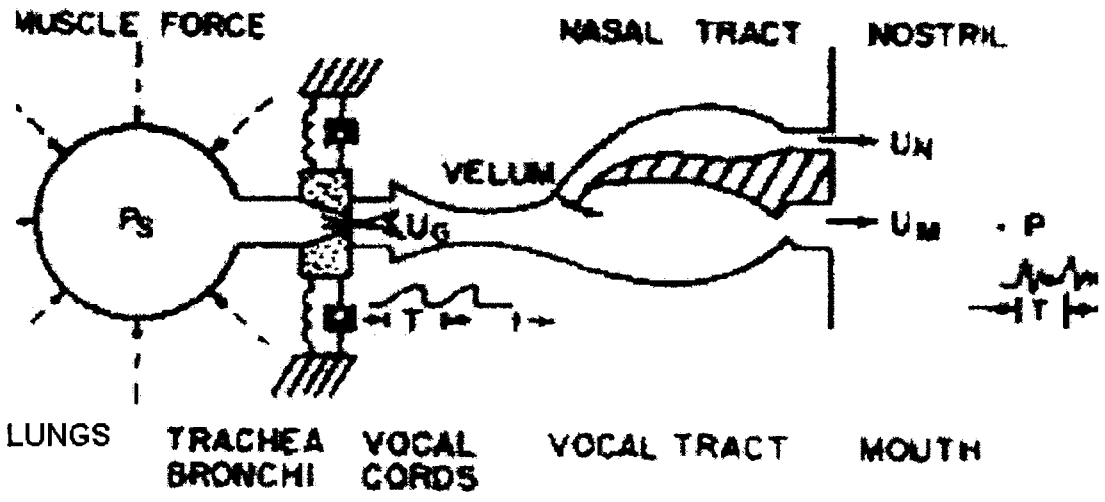

Figure 2.3.–Schematized diagram of the vocal apparatus (After Flanagan et al. [4])

The frequency spectrum is shaped by the frequencies selectivity of the tube as shown in figure 2.3. The resonance frequencies of the vocal tract tube are called formants. The formant frequencies depend upon the shape and dimension of the vocal tract. Each shape is characterized by the formants frequencies. Different sounds are performed by varying the shape of the vocal tract. Thus spectral properties of the speech signal vary with time as the vocal tract shape varies.

One of the basic method of analysis is spectrogram. The time varying spectral characteristics of speech signal can be graphically displayed through the use of sound spectrograph. It produces a 2 dimensional graph called spectrogram in which the vertical dimension corresponds to frequency and horizontal dimension to time. The darkness of the pattern is proportion to the signal energy. Thus the resonance frequencies of the vocal tract show up as dark bands in the spectrogram.

Due to the complex interplay of frequencies produce the characteristic features of sound. This emphasizes the importance of acoustic analysis of human voice or human disorders. For an extended introduction to the anatomy and physiology of the human voice production system see, e.g., [4] and [5].

2.3 EVIDENCE OF NONLINEAR BEHAVIOUR

There are a number of areas that show evidence of nonlinear behaviour in speech generation:

1. Vocal folds
2. Turbulent air flow
3. Non-plane wave propagation
4. Higher order statistics
5. Chaotic behaviour

The following sections look at each of these areas.

1. Vocal folds

There are a number of features about the oscillation of the vocal folds that show they are nonlinear: In a linear model the output is proportional to the input and yet the waveform generated from vocal fold oscillation actually changes shape under different amplitude levels. Detailed studies [6, 7] of this effect show that not only does the spectral content of the pulse alter with amplitude but also that the spectral envelope changes with fundamental frequency. The vocal chords display bifurcations [8–10]. The clearest example of this is the passage from unvoiced to voiced speech where the oscillations move from an equilibrium state, i.e. not moving, to a pseudo-periodic motion. Bifurcations are a trait of nonlinear systems but in this case there has to be a question as to whether the bifurcations are caused by the driving force passing a threshold, as in classic bifurcation systems, or whether there is some higher muscular force that is controlling the transition. Models such as multiple masses routinely include nonlinear coupling between the mass elements [11, 10]. This is based on the knowledge that the cartilage and flesh constructing the larynx have nonlinear stretching qualities. A number of works have suggested that chaotic modes of operation can be found for vocal fold

oscillation. These works should be viewed very carefully since chaos is an extremely difficult phenomenon to quantify, as will become clear throughout this thesis, and can very often cause very misleading results.

2. Turbulence

When unvoiced speech is created there is a point of constriction in the vocal tract which causes turbulence to occur. Turbulence is a nonlinear effect which occurs because of an interaction between the air flow and the acoustic sound field. As already discussed there is plenty of evidence to show that cavitation noise, which is turbulence, is chaotic and there are a number of works that suggest the same is true of fricative sounds.

3. Non-plane wave propagation

The usual model of the vocal tract is that the sound travels along the tract as plane wave propagation. Recently this view has been challenged by Teager and Teager [12] who suggest that the flow consists of a number of vortices. This work is based on examining the air flow at a range of points within the vocal tract using hot wire anemometers. If this is indeed the case then it throws into question the whole acoustic model which is based on the idea that the vocal tract can be considered as a number of acoustic tubes which have well defined reflection and standing wave properties. A good example that shows this effect is given by Kubin [13]: what is the mechanism for human whistling since no part of the vocal tract is in oscillation? The explanation given is, in summary, that an unstable jet of air is created which gives rise to vortices, when the travelling time through the vocal tract matches the frequency of the vortices then periodic vortex shedding occurs at the lips giving rise to the narrow band whistle.

4. Higher order statistics

Higher order statistics (HOS) can be used to identify the underlying nonlinearities present in a system. Unfortunately the application of HOS theories to noisy signals is very difficult and consequently the application of HOS to speech has not produced conclusive results. However what results have been published [14, 15] suggest that there is strong evidence of quadratic phase coupling, which would indicate nonlinearity.

5. Chaotic behavior

Several times in this chapter the possible existence of chaotic behaviour has been suggested. This section gives a quick overview of the work that has been conducted and some discussion of the possible shortcomings that may give rise to a number of misleading results. Most of the work in this field seems to have been inspired by the work of Teager and Teager [12] which gave clear indications that speech was nonlinear; if it is nonlinear then could it be chaotic or fractal in nature? Maragos [16] suggests that fricatives have a fractal dimension of as low as 1.7 whilst vowels may have a fractal dimension of nearer 1.2. The calculation of this dimension is through the box counting technique [17] which is restricted to a 2 dimensional plane and explains why the figures are so low, and in disparity with the dimension measures given by other authors. It should be noted that this is not saying that the measurements are wrong it is merely pointing out that the box counting dimension looks at the dimension of a waveform not of the generating system itself. The paper also shows calculation of dimension using very small data sets and showing no form of noise cancellation, these are shortcomings that are consistent with many other papers in the field. Both Boshoff [18] and McDowell and Datta [19] give similar analyses suggesting box counting dimensions of between 1 and 2 although McDowell and Datta [19] point out that the accuracy of these results is questionable. Pickover and Khorasani [20] attempt a similar analysis but on full sentences. This raises the spectre of stationarity; speech is constructed from many small segments that individually may be viewed as stationary, the normal size of these sections is about 10ms which is based on the relatively slow movement of the articulators, but a complete sentence includes many different modes of operation and indeed periods of complete silence. As a diagnostic tool this approach may have some use if it is used to compare the characteristics of different speakers saying the same sentences, but should not be used to give a definition of the fractal dimension of speech as a whole.

Marcato and Mumolo [21] show that fractal theory can be applied to the LPC to give an efficient coding of the residual signal. Fractals are similarly applied to image coding [22] and speech recognition [23].

McLaughlin and Lowry [24] use the correlation dimension to investigate a range of vowels with the conclusion that although they do seem to show low dimensional properties, the correlation dimension fails to give an accurate measure. These results are consistent with the general disenchantment with correlation dimension when applied to real world signals. Tishby [25] again examines the correlation dimension giving similar vague reports of dimensions ranging from 3 to 5 for voiced speech. He also looks at the possibility of forming a local nonlinear predictor using neural networks to enhance current predictor based systems. A similar work by Moakes and Beet [26, 27] suggests that speech is low dimensional and they apply Radial Basis Functions (RBF) to both recognition and predictive problems. Berhard and Kubin [28, 29] give preliminary evidence for low dimensional behaviour, of the order of 1 to 2, for vowels. In a very recent paper Narayanan and Alwan [30] look at fricatives showing the difficulties of convergence for the correlation dimension but suggesting low dimensions for vowels and high, around 4 to 7, dimensions for fricatives. They also examine the Lyapunov spectra suggesting that vowels have a non-chaotic structure whilst fricatives may have a single positive exponent. Bohez, Senevirathne and VanWinden [31] give a very clear application of fractal theory to recognition of vowels. Again they do not attempt to infer the actual underlying system's dimension from the fractal dimension but rather use it as a discriminatory tool. In another paper by the same authors they present an analysis of speech using an alternative box counting technique called the amplitude-scale method. Unfortunately this technique seems to give wildly different results from the box counting technique and again is limited to a 2 dimensional space. Townshend [32] gives a very full overview of the possible uses of nonlinear predictors in speech along with presenting correlation dimension results of just less than 3. These results again do not appear to be for stationary segments of speech and must be considered with care. As should be clear there has been considerable work presented in the field although on the whole the problems of noise contamination, data set size and stationarity have not been addressed fully.

2.4 REFERENCES:

1. J. Van den Berg, Myoelastic -aerodynamic theory of voice production, Journal of Speech and Hearing Reasearch , 111 2227-44, 1958 .
2. I. R. Titze, Principles of voice production, Prentice-Hall, 1994a.
3. A. Breen, Speech synthesis models: a review, Electronics and Communication Engineering Journal, pp. 19–31, February 1992.
4. J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Schafer, and N. Umeda, Synthetic Voices for Computers, IEEE Spectrum, Vo. 7, No. 10, pp. 22-45, October 1970.
5. J. Wendler, W. Seidner, G. Kittel, and U. Eysholdt. Lehrbuch der Phoniatrie und P¨adaudiologie. Georg Thieme Verlag, Stuttgart, New York, 1996.
6. I. R. Titze, The physics of small-amplitude oscillation of the vocal folds, J Acoust Soc Am, vol. 83, pp. 1536–1552, April 1988.
7. I. R. Titze, Phonation threshold pressure: A missing link in glottal aerodynamics, J Acoust Soc Am, vol. 91, pp. 2926–2935, May 1992.
8. J. Awrejcew, Bifurcation portrait of human vocal oscillations, Journal of Sound and Vibration, vol. 136, no. 1, pp. 151–156, 1990.
9. H. Herzel and J. Wendler, Evidence of chaos in phonatory samples, in EUROSPEECH 91̂, vol. 1, pp. 263–266, 1991.
10. I. Steinecke and H. Herzel, Bifurcations in an asymmetric vocal-fold model, Acoustical Society of America, vol. 97, pp. 1874–1884, March 1995.
11. T. Koizumi, S. Taniguchi, and S. Hiromitsu, Two-mass models of the vocal cords for natural sounding voice synthesis, Journal of the Acoustical Society of America, vol. 8, pp. 1179–1192, October 1987.

12. H. M. Teager and S. M. Teager, Evidence for nonlinear sound production mechanisms in the vocal tract, in Proc NATO ASI on Speech Production and Speech Modelling, pp. 241–261, 1990.

13. G. Kubin, Nonlinear processing of speech, in Speech Coding and Synthesis (W. B. Kleijn and K. K. Paliwal, eds.), pp. 557–610, Amsterdam: Elsevier, 1995.

14. S. McLaughlin, S. Hovel, and A. Lowry, Identification of nonlinearities in vowel generation, in EUSIPCO 94, vol. 2, pp. 1133–1137, Elsevier Science, 1994.

15. J. W. A. Fackrell and S. McLaughlin, The higher order statistics of speech signals, in IEE Colloquium on Techniques for speech processing and their application, no. 1994/138, pp. 7/1–7/6, IEE, June 1994.

16. P. Maragos, Fractal aspects of speech signals: dimension and interpolation, in ICASSP 91, pp. 417–420, IEEE, 1991.

17. L. S. Liebovitch and T. Toth, A fast algorithm to determine fractal dimension by box counting, Physics Letters A, vol. 141, pp. 386–390, November 1989.

18. H. F. V. Boshoff and M. Grotepass, The fractal dimension of fricative speech sounds," in COSMIG '91, pp. 12–16, IEEE, 1991.

19. P. S. McDowell and S. Datta, The fractal characterisation of isolated human speech, Proceedings of the Institute of Acoustics, vol. 16, no. 5, pp. 247–253, 1994

20. C. A. Pickover and A. Khorasani, Fractal characterization of speech waveform graphs, Comput and Graphics, vol. 10, no. 1, pp. 51–61, 1986.

21. L. Marcato and E. Mumolo, Coding of speech signal by fractal techniques, in EUROSPEECH '93, pp. 745–748, 1993.

22. M. F. Barnsley and A. D. Sloan, A better way to compress images, BYTE, pp. 215–223, January 1989.

23. E. L. J. Bohez, T. R. Senevirathne, and J. A. VanWinden, Fractal dimension and iterated function system (ifs) for speech recognition, Electronic Letters, vol. 28, pp. 1382–1384, July 1992.

24. S. McLaughlin and A. Lowry, Nonlinear dynamical systems concepts in speech analysis, in EUROSPEECH '93, pp. 377–380, 1993.

25. N. Tishby, A dynamical systems approach to speech processing, in ICASSP '90, pp. 365–368, IEEE, 1990.

26. P. A. Moakes and S.W. Beet, Recurrent radial basis functions for speech period detection, Proceedings of the Institute of Acoustics, vol. 16, no. 5, pp. 271–278, 1994.

27. P. A. Moakes and S.W. Beet, Analysis of non-linear generating dynamics, in ICSLP 94, pp. 1039–1042, 1994.

28. H. P. Bernhard and G. Kubin, Detection of chaotic behaviour in speech signals using fraser's mutual information algorithm, in 13th GRETSI symposium on signal and image processing, 1991.

29. H. P. Bernhard and G. Kubin, Speech production and chaos, in 13th GRETSI symposium on signal and image processing, 1991.

30. S. S. Narayanan and A. A. Alwan, A nonlinear dynamical system analysis of fricative consonants, The Journal of the Acoustical Society of America, vol. 97, pp. 2511–2524, April 1995.

31. T. R. Senvirathne, E. L. J. Bohez, and J. A. VanWinden, Amplitude scale method: Newand efficient approach to measure fractal dimension of speech waveforms, Electronics Letters, vol. 28, pp. 420–422, February 1992.

32. B. Townshend, Nonlinear prediction of speech, in IACSSP '91, pp. 425–428, IEEE, 1991.