3

Mathematical Modeling of Speech Production and Its Application to Noise Cancellation

N. R. Raajan[1], T. R. Sivaramakrishnan[1] and Y. Venkatramani[2]
[1]*School of Electrical and Elcetronics Engineering, SASTRA University, Thanjore*
[2]*Saranathan College of Engineering, Trichy*
India

1. Introduction

Sound emanates by three processes, they are twisting of nerves, wires beating of membranes or blowing of air through holes. But human voice mechanism is different as it comes out in different languages and feelings by a control mechanism, the brain. As per Indian thought The soul (Atma) associates with (budhi) brain , and later inturn orders the (manas) heart. Thus the (manas) heart under the influence of (bhudhi) brain stimulates the (jathagani) simulator. The Jathagani stimulates Udanda vata and finally the (intuition) udana vata produces speech. The voice with which we speak has two components namely **1)Dhwanyaatmaka sabdas 2)VarNaatmaka sabdas**.

Dhwanyaatmaka sabdas (fricative sound)are produced as sounds without modification. These sounds are modified after they come out of the vocal cords into pharynx and mouth. Here by different types of movements in pharynx, palate, tounge checks and lips, various syllables and words are produced. the production of speech will be effected by the action of the areas of cerebral cortex viz, 1) Audio sensory, 2) Audio Psychic and 3) Audio-motor. Simply, **Dhvanyaatmaka** (fricative sound), for example, is the sound produced by the beat of a drum or the ringing of a bell, etc. **VarNaatmaka** (Plosive sound) is the sound produced by the vocal organs, namely, the throat, palate etc. For example, the sound of the letter, ka, kha, etc.

dhvani visheSasahakrta kanThataalva |
bhighaata janyashca varNaatmaka ||
shabdaartha Ratnaakara ||

Block diagram shows the complete process of producing and perceiving speech from the formulation of a message in the brain of a talker, to the creation of the speech signal, and finally to the understanding of the message by a listener. In their classic introduction to speech science. The process starts in the upper left as a message represented somehow in the brain of the speaker. The message information can be thought of as having a number of different representations during the process of speech production.

For example the message could be represented initially as English text. In order to *speak* the message, the talker implicitly converts the text into a symbolic representation of the sequence

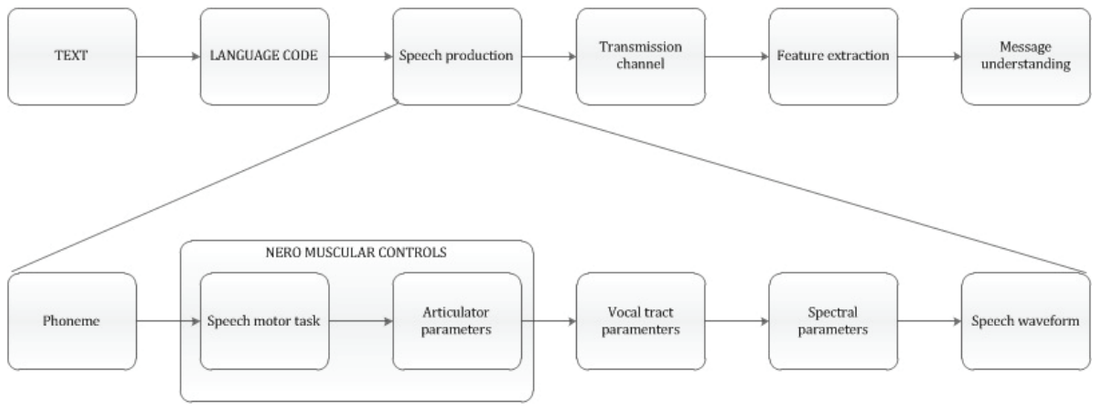


Fig. 1. Block Diagram

of sounds corresponding to the spoken version of the text. This step, called the language code generator (it is done under bhudhi (brain) converts text to speech)in Block digaram, converts text symbols to phonetic symbols (along with stress and durational information) that describe the basic sounds of a spoken version of the message and the manner (i.e., the speed and emphasis) in which the sounds are intended to be produced. As there are labeled with phonetic symbols using a computer-keyboard-friendly code called ARPAbet. Thus, the text *shouldwechase* is represented phonetically (in ARPAbet symbols) as [SH UH D W IY CH EY S]. The third step in the speech production process is the conversion to *neuromuscularcontrols*, i.e., the set of control signals that direct the neuromuscular system to move the speech articulators, namely the tongue, lips, teeth,jaw and velum, in a manner that is consistent with the sounds of the desired spoken message and with the desired degree of emphasis. The end result of the neuromuscular controls step is a set of articulatory motions (continuous control) that cause the vocal tract articulators to move in a prescribed manner in order to create the desired sounds. Finally the last step in the Speech Production process is the *vocaltractsystem* that physically creates the necessary sound sources and the appropriate vocal tract shapes over time so as to create an acoustic waveform, that encodes the information in the desired message into the speech signal. To determine the rate of information flow during speech production, assume that there are about 32 symbols (letters) in the language(in English there are 26 letters, but if we include simple punctuation. we get a count closer to 32 = 2^5symbols). Furthermore, the rate of speaking for most people is about 10 symbols per second (somewhat on the high side, but still acceptable for a rough information rate estimate). Hence, assuming independent letters as a simple approximation, we estimate the base information rate of the text message as about 50 bps (5 bits per symbol times 10 symbols per second). At the second stage of the process, where the text representation is converted into phonemes and prosody (e.g., pitch and stress) markers, the information rate is estimated to increase by a factor of 4 to about 200 bps. For example, the ARBAbet phonetic symbol set used to label the speech sounds contains approximately 64 = 2^6 symbols, or about 6 bits/phoneme (again a rough approximation assuming independence of phonemes). There are 8 phonemes in approximately 600ms. This leads to an estimate of 8 Œ 6/0.6 = 80 bps. Additional information required to describe prosodic features of the signal (e.g., duration, pitch, loudness) could easily add 100 bps to the total information rate for a message encoded as a speech signal.

The information representations for the first two stages in the speech signal are discrete so we can readily estimate the rate of information flow with some simple assumptions. For the next stage in the speech production part of the speech chain, the representation becomes continuous (in the form of control signals for articulatory motion). If they could be measured, we could estimate the spectral bandwidth of these control signals and appropriately sample and quantize these signals to obtain equivalent digital signals for which the data rate could be estimated. The articulators move relatively slowly compared to the time variation of the resulting acoustic waveform. Estimates of bandwidth and required accuracy suggest that the total data rate of the sampled articulatory control signals is about 2000 bps. Thus, the original text message is represented by a set of continuously varying signals whose digital representation requires a much higher data rate than the information rate that we estimated for transmission of the message as a speech signal.

Finally, as we will see later, the data rate of the digitized speech waveform at the end of the speech production part of the speech chain can be anywhere from 64,000 to more than 700,000 bps. We arrive at such numbers by examining the sampling rate and quantization required to represent the speech signal with a desired perceptual fidelity. For example, *telephonequality* requires that a bandwidth of 0 to 4 kHz be preserved, implying a sampling rate of 8000 samples/sec. Each sample can be quantized with 8 bits on a log scale, resulting in a bit rate of 64,000 bps. This representation is highly intelligible (i.e., humans can readily extract the message from it) but to most listeners, it will sound different from the original speech signal uttered by the talker. On the other hand, the speech waveform can be represented with *CDquality* using a sampling rate of 44,100 samples/s with 16 bit samples, or a data rate of 705,600 bps. In this case, the reproduced acoustic signal will be virtually indistinguishable from the original speech signal. As we move from text to speech waveform through the speech chain, the result is an encoding of the message that can be effectively transmitted by acoustic wave propagation and robustly decoded by the hearing mechanism of a listener. The above analysis of data rates shows that as we move from text to sampled speech waveform, the data rate can increase by a factor of 10,000. Part of this extra information represents characteristics of the talker such as emotional state, speech mannerisms, accent, etc., but much of it is due to the inefficiency of simply sampling and finely quantizing analog signals. Thus, motivated by an awareness of the low intrinsic information rate of speech, a central theme of much of digital speech processing is to obtain a digital representation with lower data rate than that of the sampled waveform.

One of the features which has bothered researchers in the area of speech synthesis in the past has been voicing. We discuss this here because it is a good example of how failure to understand the differences between abstract and physical modeling can lead to disproportionate problems (Keating 1984). The difficulty has arisen because of the nonlinearity of the correlation between the cognitive phonological voicing and how the feature is rendered phonetically. Phonological voicing is a distinctive feature in that, it is a parameter of phonological segments the presence or absence of which is able to change one underlying segment into another. For example, the English alveolar stop /d/ is [+voice] (has voicing) and differs on this feature from the alveolar stop / t / which is [-voice] (does not have voicing). Like all phonological different features, the representation is binary, meaning in this case that [voice] is either present or absent in any one segment.

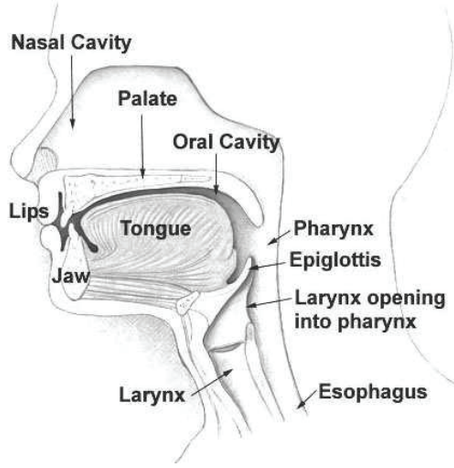


Fig. 2. Anatomy of vocal fold

The most frequent phonetic parameter to correlate with phonological voicing is vocal cord vibration. The vocal cords usually vibrate when the underlying plan is to produce a [+voice] sound, but usually do not when the underlying plan is to produce a [-voice] sound.

Many synthesis models assume constant voicing vocal-cord vibration, but it is quite clear that the binary distinction of vocal-cord vibration vs. no vocal-cord vibration is not accurate. Vocal-cord vibration can begin abruptly (as when there is a glottal stop onset to make this possible singers regularly do this), gradually (the usual case), or at some point during the phone, although it may be phonologically voiced. Similarly for phonologically voiceless segments, it is certainly not the case that on every occasion there is no vocal-cord vibration present at some point during the phone. We know of no model which sets out the conditions under which these variants occur.

Phonological characterizations of segments should not be considered as though they were phonetic, and sets of acoustical features should not be given one-to-one correlation with phonological features. More often than not the correlation is not linear nor, apparently, consistent- though it may yet turn out to be consistent in some respects. Phonology and phonetics cannot be linked simply by using phonological terms within the phonetic domain such as the common transfer of the term voicing between the two levels. Abstract voicing is very different from physical voicing, which is why we consistently use different terms for the two. The basis of the terminology is different for the two levels; and it is bad science to equate the two so directly.

Major problem in speech processing is to represent the shape and characteristics of the vocal tracts. This task is normally done by using an acoustics tube model, based on the calculation of the area function. A Mathematical model of Vocal fold has been obtained as part of new approach for Noise cancellation.

2. The physics of sound production

Speech is the unique signal generated by the human vocal apparatus. Air from the lungs is forced through the vocal tract, generating acoustic waves that are radiated from the lips as a

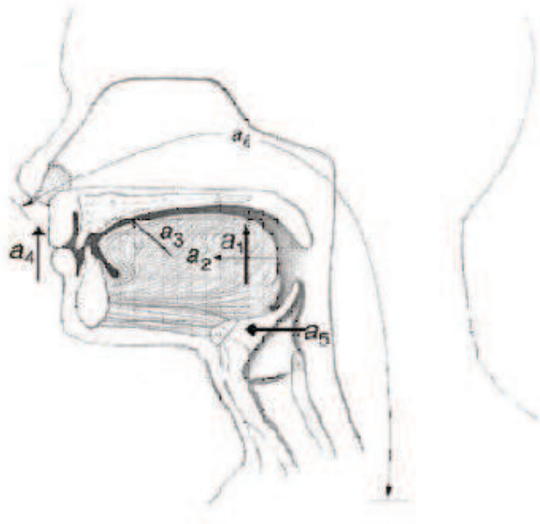


Fig. 3. Articulatory model

pressure field. The physics of this process is well understood, giving us important insights into (sound) speech communication.

The rudiments of speech generation are given in next two sections. Thorough treatments of this important subject can be found in [Flanagan] and [Rabiner and Schafer].

2.1 Development of speech

An young child for the first few months of his life goes on hearing the words being spoken by the persons around him, suppose he has heard the word 'AMMA' serval times spoken by his parents etc., Then he goes on thinking about the production of that word with his audio psychic area ties to reproduce with different movements of his lips, tongue etc., This will be effected by his audio-motor area thus after a few trials the child will be able to reproduce that word. the speech is nothing but a modified expiratory act produced while the expiratory air vibrates the vocal cords of the larynx, and altered by the movements of different structures like tongue, lips, etc.

2.2 The human vocal apparatus

Fig:2 shows a representation of the mid sagittal section of the human vocal tract [Coker]. In this model, the cross-sectional area of the oral cavity A(x), from the glottis, x = 0, to the lips, x = L, is determined by five parameters: a_1 , tongue body height; a_2 , anterior/posterior position of the tongue body; a_3 , tongue tip height; a_4 , mouth opening; and a_5 , pharyngeal opening. In addition, a sixth parameter, a_6 , is used to additively alter the nominal 17-cm vocal tract length. The articulatory vector a is $(a_1, a_2, ..., a_6)$.

The vocal tract model has three components: an oral cavity, a glottal source, and an acoustic impedance at the lips. We shall consider them singly first and then in combination. As is commonly done, we assume that the behavior of the oral cavity is that of a lossless acoustic

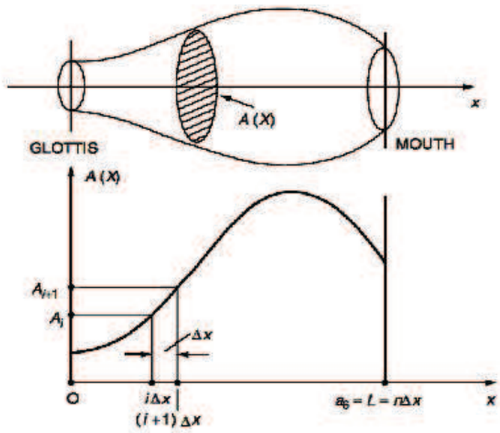


Fig. 4. The acoustic tube model of the vocal tract and its area function

tube of slowly varying in time and space cross-sectional area, A(x), in which plane waves propagate in one dimension (see Fig:3). [Sondhi] and [Portnoff] have shown that under these assumptions, the pressure, p(x, t), and volume velocity, u(x, t), satisfy

The vocal tract model has three components: an oral cavity, a glottal source, and an acoustic impedance at the lips. We shall consider them singly first and then in combination. As is commonly done, we assume that the behavior of the oral cavity is that of a lossless acoustic tube of slowly varying (in time and space) cross-sectional area, A(x), in which plane waves propagate in one dimension. [Sondhi] and [Portnoff] have shown that under these assumptions, the pressure, p(x, t), and volume velocity, u(x, t), satisfy

$$-\frac{\partial p}{\partial x} = \frac{\rho}{A(x,t)}\frac{\partial u}{\partial t} \tag{1}$$

and

$$-\frac{\partial u}{\partial x} = \frac{A(x,t)}{\rho c^2}\frac{\partial p}{\partial t} \tag{2}$$

which express Newton's law and conservation of mass, respectively. In above stated equation ρ is the equilibrium density of the air in the tube and c is the corresponding velocity of sound. Differentiating (eq:1) and (eq:2) with respect to time and space, respectively, and then eliminating the mixed partials, we get the well-known Webster equation [Webster] for pressure,

$$\frac{\partial^2 p}{\partial x^2} + \frac{1}{A(x,t)}\frac{\partial p \partial A}{\partial x \partial x} = \frac{1\partial^2 p}{c^2 \partial t^2} \tag{3}$$

The eigenvalues of (eq:3) are taken as formant frequencies. It is preferable to use the Webster equation (in volume velocity) to compute a sinusoidal steady-state transfer function for the acoustic tube including the effects of thermal, viscous, and wall losses.

To do so we let $p(x,t) = P(x,\omega)$ and $u(x,t) = U(x,\omega)$, where ω is angular frequency. When p and u have this form, (eq:1) and (eq:2) become (cf. [Rabiner, L.R.and Schafer, R.W.]) and

$$p(x,t) = P(x,\omega) \tag{4}$$

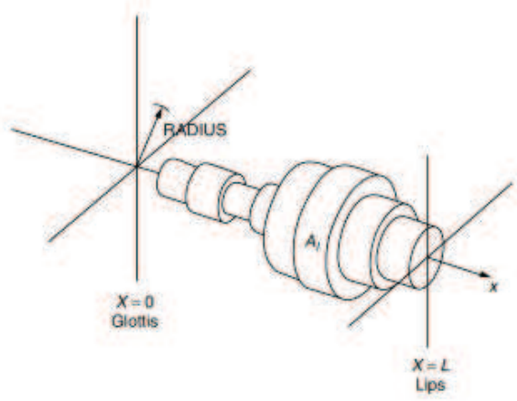


Fig. 5. The discretized acoustic tube model of the vocal tract

$$u(x,t) = U(x,\omega) \tag{5}$$

respectively. In order to account for the losses we define $Z(x,\omega)$ and $Y(x,\omega)$ to be the generalized acoustic impedance and admittance per unit length, respectively. Differentiating (eq:5) with respect to x and substituting for $\frac{-dp}{dx}$ and P from (eq:4) and (eq:5), respectively, we obtain

$$\frac{d^2U}{dx^2} = \frac{1}{Y(x,\omega)}\frac{dU}{dx}\frac{dY}{dx} - Y(x,\omega)Z(x,\omega)U(x,\omega) \tag{6}$$

This is recognized as the *"lossy"* Webster equation for volume velocity. The sinusoidal steady-state transfer function of the vocal tract can be computed by discretizing (eq:6) in space and obtaining approximate solutions to the resulting difference equation for a sequence of frequencies. Let us write U_i^k to signify $U(i\Delta x, k\Delta\omega)$ where the spatial discretization assumes $\Delta x = \frac{L}{n}$ with i = 0 at the glottis and i = n at the lips, as is shown in Fig:4. Similarly, we choose $\Delta\omega = \frac{\Omega}{N}$ and let $0 \leq$ k $\leq$ N . We shall define A_i, Y_i^k, and Z_i^k in an analogous manner. Approximating second derivatives by second central differences and first derivatives by first backward differences, the finite difference representation of (eq:6) is given by (eq:7)

$$U_{i+1}^k = U_i^k\left(3 + (\Delta x)^2 Z_i^k Y_i^k - \frac{Y_{i-1}^k}{Y_i^k}\right) + U_{i-1}^k\left(\frac{Y_{i-1}^k}{Y_i^k} - 2\right) \tag{7}$$

Given suitable values for U_0^k and U_1^k for $0 \leq k \leq N$, we can obtain the desired transfer functions from (eq:7). We must find appropriate expressions for Y and Z to account for the losses. Losses arise from thermal effects and viscosity and primarily due to wall vibrations. A detailed treatment of the wall losses is found in [Portnoff] and is summarized by [Rabiner and Schafer]. Portnoff assumes that the walls are displaced $\xi(x,t)$ in a direction normal to the flow due to the pressure at x only. The vocal tract walls are modeled by a damped spring-mass system for which the relationship between pressure and displacement is

$$p(x,y) = M\frac{\partial^2\xi}{\partial t^2} + b\frac{\partial\xi}{\partial t} + k(x)\xi(x,t) \tag{8}$$

where M , b, and k(x) are the unit length wall mass, damping coefficient, and spring constant, respectively. The displacement of the walls is assumed to perturb the area function about a

neutral position according to

$$A(x,t) = A(x) + S(x)\xi(x,t) \tag{9}$$

where A(x) and S(x) are the neutral area and circumference, respectively. By substituting (eq:1) into (eq:2), ignoring higher-order terms, transforming into the frequency domain, Portnoff goes on to observe that the effect of vibrating walls is to add a term to the acoustic admittance in (eq:5), where

$$Y_w(x,t) = j\varpi S(x,\omega)\left(\frac{[k(x)-\omega^2 M]-j\omega b}{[k(x)-\omega^2 M]^2+\omega^2 b^2}\right) \tag{10}$$

The other losses that we wish to consider are those arising from viscous friction and thermal conduction. The former can be accounted for by adding a real quantity Z_v to the acoustic impedance in (eq:4),

$$Z_v(x,\omega) = \frac{S(x)}{A^2(x)}\left(\frac{\omega\rho\mu}{2}\right)^{\frac{1}{2}} \tag{11}$$

Here μ is the viscosity of air. The thermal losses have an effect which is described by adding a real quantity Y_T to the acoustic admittance in (eq:5), where

$$Y_T(x,\omega) = \frac{S(x)(\eta-1}{\rho c^2}\left(\frac{\lambda\omega}{2C_p\rho}\right)^{\frac{1}{2}} \tag{12}$$

Here A is the coefficient of heat conduction, η is the adiabatic constant, and C_p is the heat capacity. All the constants are, of course, for air at the conditions of temperature, pressure, and humidity found in the vocal tract. In view of (eq:1), (eq:2), (eq:10), (eq:11) and (eq:12) it is possible to set

$$Z(x,\omega) = \frac{j\omega\rho}{A(x)+Z_v(x,\omega)} \tag{13}$$

and

$$Y(x,\omega) = \frac{j\omega A(x)}{\rho c^2} + Y_w(x,\omega) + Y_T(x,\omega) \tag{14}$$

There are two disadvantages to this approach. First, (eq:13) and (eq:14) are computationally expensive to evaluate. Second, (eq:10) requires values for some physical constants (saliva, phlegm, tonsils, etc.,) of the tissue forming the vocal tract walls. Estimates of these constants are available in [Rabiner, L.R.and Schafer, R.W.] and [Webster]. A computationally simpler empirical model of the losses which agrees with the measurements has been proposed by [Sondhi] in which

$$Z(x,\omega) = \frac{j\omega\rho}{A(x)} \tag{15}$$

and

$$Y(x,\omega) = \frac{A(x)}{\rho c^2}\left(j\omega + \frac{\omega_0^2}{\alpha+j\omega} + (\beta j\omega)\right)^{\frac{1}{2}} \tag{16}$$

Sondhi [10] has chosen values for the constants, w_0 = 406π, α = 130π, β = 4, which he then shows give good agreement with measured formant bandwidths. Moreover, the form of the model agrees with the results of Portnoff, which becomes clear when we observe that $Y_w(x,\omega)$ in (eq:10) will have the same form as the second term on the right-hand side of (eq:16) if

k(x)≡0 and the ratio of circumference to area is constant. In fact, Portnoff used k(x) = 0 and this assumption is reasonable. The third term in the right-hand side of (eq:16) is of the same form as (eq:11) and (eq:12) (under the assumption that the ratio of S to A is constant) by noting that

$$(j\omega)^{\frac{1}{2}} = (1+j)\left((\frac{\omega}{2}\right)^{\frac{1}{2}} \tag{17}$$

2.3 Boundary conditions

With a description of the vocal tract in hand, we can turn our attention to the boundary conditions. Following [Flanagan], the glottal excitation has been assumed to be a constant volume source with an asymmetric triangular waveform of amplitude V . [Dunn et al.] have analyzed such a source in detail. What is relevant is that the spectral envelope decreases as the square of frequency. We have therefore taken the glottal source $U_g(\omega)$ to be

$$U_g(\omega) = \frac{V}{\omega^2} \tag{18}$$

For the boundary condition at the mouth, the well-known [Portnoff] and [Rabiner and Schafer] relationship between sinusoidal steady-state pressure and volume velocity, is used.

$$P(L,\omega) = Z_r(\omega)U(L,\omega) \tag{19}$$

Here the radiation impedance Z_r is taken as that of a piston in an infinite plane baffle, the behavior of which is well approximated by

$$Z_r(\omega) = \frac{j\omega L_r}{\left(\frac{1+j\omega L_r}{R}\right)} \tag{20}$$

Values of the constants which are appropriate for the vocal tract model are given by [Flanagan] as

$$R = \frac{128}{9\pi^2} \tag{21}$$

and

$$L_r = 8\left[A(L)/\pi\right]^{\frac{1}{2}}/3\pi c \tag{22}$$

It is convenient to solve (eq:6) with its boundary conditions (eq:19) and (eq:20) by solving a related initial-value problem for the transfer function

$$H(\omega) = U(L,\omega)/U(0,\omega) \tag{23}$$

$$-\frac{dU}{dx}|_{x=L} = \frac{A(L)}{\rho c^2}(j\omega)P(L,\omega) \tag{24}$$

From which the frequency domain difference equation is

$$-\frac{U_n^k - U_{n-1}^k}{\Delta x} = jk\Delta\omega\frac{A_n}{\rho c^2}P_n^k \tag{25}$$

been derived. Let it be noted from (eq:21), finally, the vocal track output is obtained by (eq:26)

$$\xi(x,t) = \mathrm{e}^{1/2\,\frac{\left(-b+\sqrt{b^2-4k(x)M}\right)t}{M}} + \mathrm{e}^{-1/2\,\frac{\left(b+\sqrt{b^2-4k(x)M}\right)t}{M}} - \frac{1}{\sqrt{b^2-4\,k\,(x)\,M}} - \left(\int p\,(x,t)\,\mathrm{e}^{1/2\,\frac{\left(b+\sqrt{b^2-4k(x)M}\right)t}{M}}\,dt\right)\mathrm{e}^{-1/2\,\frac{\left(b+\sqrt{b^2-4k(x)M}\right)t}{M}} + \left(\int p\,(x,t)\,\mathrm{e}^{-1/2\,\frac{\left(-b+\sqrt{b^2-4k(x)M}\right)t}{M}}\,dt\right)\mathrm{e}^{1/2\,\frac{\left(-b+\sqrt{b^2-4k(x)M}\right)t}{M}}\left(\mathrm{e}^{-\frac{bt}{M}}\right) \tag{26}$$

Here, p represents pressure, $k(x)$ - Damping coefficient, M - Mass of speech, $\xi(x,t)$ - resultant Value. [NRR]

3. Non- stationary speech signal

The speech signal is the solution to equation(eq:3) Since the function A(x,t) is continuously varyinging time,the solution, p(t),is a non-stationary random change in time. Fortunately, A(x,t) is slowly time-varying with respect to p(t).That is,

$$\overline{P}_k = \overline{P}(k\Delta\omega) = H(k\Delta\omega)U_g(k\Delta\omega)Z_r(k\Delta\omega) \tag{27}$$

$$\left|\frac{\partial A}{\partial t}\right| << \left|\frac{\partial p}{\partial t}\right| \tag{28}$$

Equation (eq:28) may be taken to mean that p(t) is quasi-stationary or piecewise stationary. As such, p(t) can be considered to be a sequence of intervals within each one of which p(t) is stationary. It is true that there are rapid articulatory gestures that violate (eq:28), but in general the quasi-stationary assumption is useful.

4. Fluid dynamical effects

Equation (eq:3) predicts the formation of planar acoustic waves as a result of air flowing into the vocal tract according to the boundary condition of (xyz). However, the Webster equation ignores any effects that the convective air flow may have on the function p(t).

If, instead of (eq:1) and (eq:2), we consider two-dimensional wave propagation, conservation of mass can be written as

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial y} = -M^2\frac{\partial p}{\partial t} \tag{29}$$

where M is the Mach number.
We can also include the viscous and convective effects by observing

$$\frac{\partial u_x}{\partial t} = -\frac{\partial p}{\partial x} - \frac{\partial(u_x u_y)}{\partial x} + \frac{\partial}{\partial x}\left[\frac{1}{N_R}\left(\frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial x}\right) - \overline{\mu_x \mu_y}\right] \tag{30}$$

$$\frac{\partial u_y}{\partial t} = -\frac{\partial p}{\partial y} - \frac{\partial(u_x u_y)}{\partial y} + \frac{\partial}{\partial y}\left[\frac{1}{N_R}\left(\frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial y}\right) - \overline{\mu_x \mu_y}\right] \tag{31}$$

In (eq:30) and (eq:31) the first term on the right-hand side is recognized as Newton's law expressed in (eq:1) and (eq:2). The second term represents the convective flow. The third term accounts for viscous shear and drag at Reynolds number, NR, and the last term represents turbulence.

Equations (eq:29),(eq:30) and (eq:31) are known as the normalized, two-dimensional, Reynolds averaged, Navier - Stokes equations for slightly compressible flow. These equations can be solved numerically for p(t). The solutions are slightly different from those obtained from (eq:3) due to the formation of vortices and transfer of energy between the convective and wave propagation components of the fluid flow. Typical solutions for the articulatory configuration of Fig: 2 are shown in eqns. (eq:5) and (eq:6). There is reason to believe that (eq:29),(eq:30) and (eq:31) provide a more faithful model of the acoustics of the vocal apparatus than the Webster equation does [11].

5. Noise cancellation

The conclusion to be drawn from the previous two sections is that information is encoded in the speech signal in its short-duration amplitude spectrum [Rabiner, L.R. and Schafer, R.W.]. This implies that by estimating the power spectrum of the speech signal as a function of time, we can identify the corresponding sequence of sounds. Because the speech signal x(t) is non-stationary it has a time-varying spectrum that can be obtained from the time-varying Fourier transform, $X_n(\omega)$. Note that x(t) is the voltage analog of the sound pressure wave, p(t), obtained by solving (eq:3).

5.1 Algorithm

1. Read b, t, M, k(x)
2. $d = b^2 - 4 * k(x) * M$
3. $z = \sqrt{d}$
4. $u = -b + z$
5. $v = b + z$
6. Get the value of f$\rightarrow$ function
7. Read lower and upper limits
8. Read n$\rightarrow$ numbr of iterations
9. h = $\frac{(upper - lower)}{2}$
10. $S = F(a)$
11. for i = 1: 2: (n-1) (odd)
12. x = a+$h. * i$
13. S = S+4 $* f(x)$
14. end
15. for i=2:2:(n-2)
16. x = a + $h. * i$

17. S = S + 2 ∗ $f(x)$
18. end
19. S = S + f(b)
20. A = h* $\frac{s}{3}$
21. Repeat the step 6 to 20
22. B = h* $\frac{s}{3}$
23. Obtained values are substuited $\xi(x,t) = exp(0.5 * u * t/M) + exp(-0.5 * v * t/M) + \frac{1}{z} * [A * exp(t * z/M) - B * exp(-0.5 * v * t/M)]$
24. Plot the sequence in the polar graph

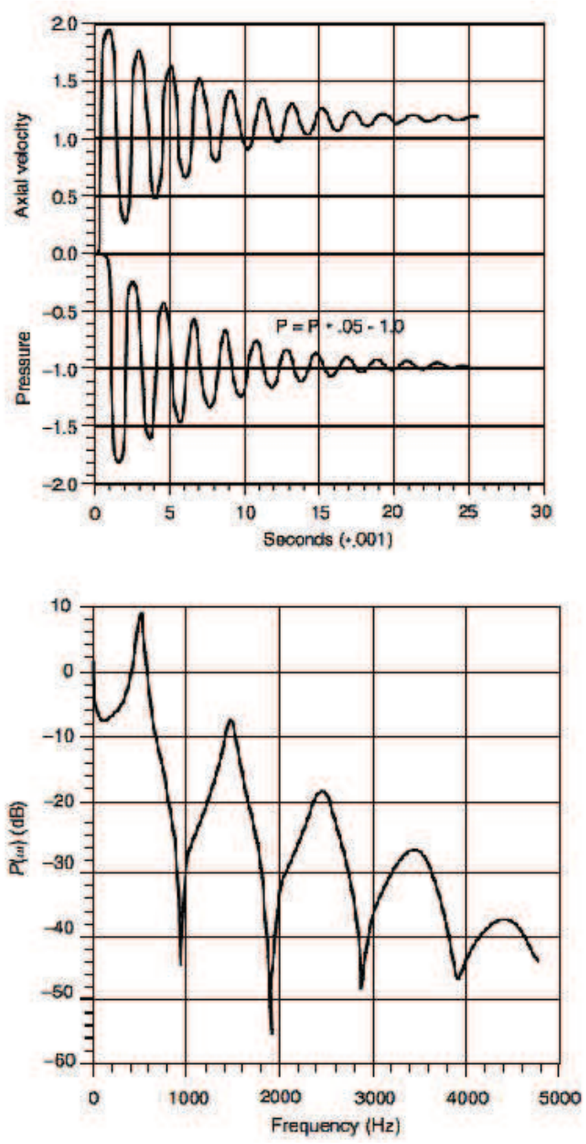


Fig. 6. Speech signals and their spectrum obtained by solving the *NavierStokes* equations

The result is obtained by implementation of the equation (eq:26) in MATLAB. The various values for the pressure, Damping co-efficient and mass is considered for the implementation of the noise cancellation. The graphs are plotted with basic, mid and high damping efficiency (Fig: 7, Fig: 8 and Fig: 9) respectively. Fig: 10 shows the original signal with noise and without noise through the equation (eq:26). From this it can be concluded that the equation can be implemented for active noise cancellation.

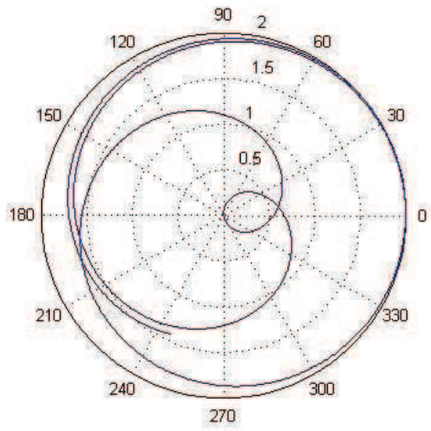


Fig. 7. The acoustic tube model of the vocal tract with basic damping efficiency

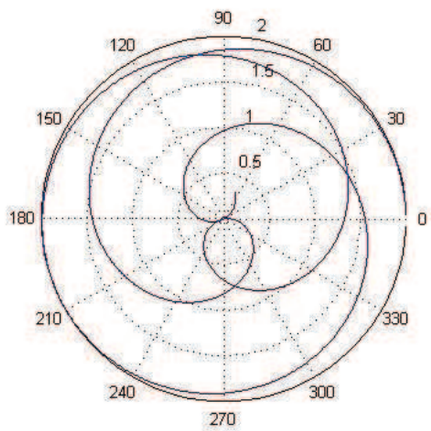


Fig. 8. The acoustic tube model of the vocal tract with Mid damping efficiency

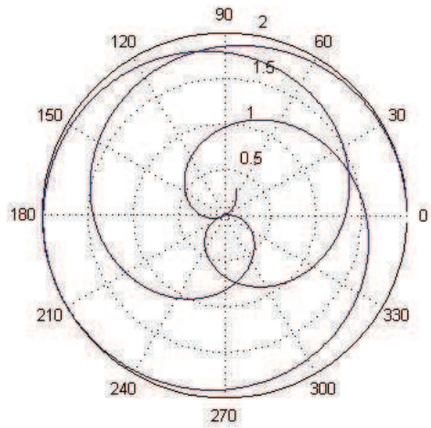


Fig. 9. The acoustic tube model of the vocal tract with high damping efficiency

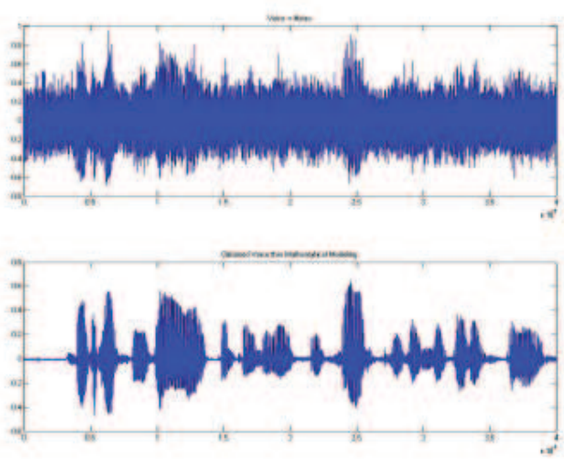


Fig. 10. (a) voice + Noise (b) Voice obtained through modeling

6. Conclusion

Producing an opposing signal (anti-noise) with the same amplitude as the noise you want to reduce (unwanted noise) but with the opposite phase, yields a significant reduction in the noise level. ANC tries to eliminate sound components by adding the exact opposite sound. The level of attenuation is highly dependent on the accuracy of the system for producing the amplitude and the phase of the reductive signal (anti-noise).

The mathematical modeling of vocal fold will recognize only the voice it never create a signal opposite to the noise. It will feed only the vocal output and not the noise, Since it uses shape and characteristic of speech.

The parameters of clean speech sample considered for testing of the algorithms were: duration 2 seconds, PCM 22.050 kHz, 8 bit mono sample recorded under laboratory conditions. This

	Female	Male
Fundamental frequency F_0 (Hz)	207	119
Glottal peak flow	0.14	0.23
Closed quotient	0.26	0.39

Table 1. Properties of the glottal wave (Normal phonation)

Configuration	Thickness lip	Length lip
Basic	0.25	7
Long lip	0.25	9
Short lip	0.25	5
Higher opening	0.25	7
Trapered	0.25(bottom) 0.125 (free tip)	7
thin lip	0.125	7

Table 2. Properties of Mouth

Background noise	Parametric background quality
High	Hissing - Fizzing
Mid	Rushing - Roaring
Low	Rumbling -Rolling
Buzz	Humming - Buzzing
Flutter	Bubbling - Percolating
Static	Crackling - Staticky

Table 3. Parametric estimationof noise

Noise	MMVF	LMS	RLS	AFA	NLMS
acoustic	28.341	23.988	18.729	22.669	20.146
Short	23.105	19.769	20.161	18.083	20.019
White	30.142	20.581	25.105	28.565	26.206
Echo and fading	26.231	21.499	19.281	20.042	26.165

Table 4. Result Obtained

speech signal is used as a benchmark for speech processing. Various noises were generated and added to the original speech signal. The SNR of the signal corrupted with the noise was 8 dB. A linear combination of the generated noise and the original signal is used as the primary input. The outputs SNR of the denoised speech signal are calculated.

MMVF - Mathematical modeling of Vocal fold
LMS - Least mean square
RLS - Recursive least square
AFA - Adaptive filter algorithm (Adaptive RLS)
NLMS - Normalized LMS

7. References

[1] N. R. Raajan, Y. Venkaaramani, & T. R. Sivaramakrishnan *A novel approach to noise cancellation for communication devices*, Instrumentation Science and Technology, Taylor and Francis group, Volume 37, Issue 6, PP: 720-729, 2009.

[2] Fitch, H. L. *Reclaiming temporal information after dynamic time warping.*, J. Acoust. Soc. Amer. 1983 , 74 (Suppl. 1), 816.

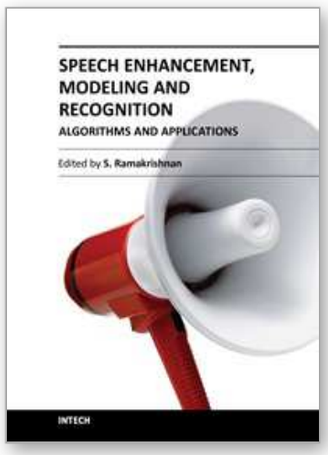
[3] Coker, C. H. *A model of articulatory dynamics and control*, Proc. IEEE, pp. 452-460, 1989.

[4] Portnoff, M. R. *A quasi-one-dimensional digital simulation for the time varying vocal tract.*,Masters thesis, MIT, 1973.

[5] Rabiner, L. R. & Schafer, R. W. *Digital Processing of Speech Signals*, Prentice Hall: Englewood Cliffs, NJ, 1978.

[6] Sondhi, M. M. *Model for wave propagation in a lossy vocal tract.*, PP. 1070 - 1075, J. Acoust. Soc. Amer. 1974, PP:55 - 67.

[7] Webster, A. G. *Acoustical impedance and the theory of horns.*, Proc. Nat. Acad. Sci. 1919, PP. 275-282.



Speech Enhancement, Modeling and Recognition- Algorithms and Applications

Edited by Dr. S Ramakrishnan

ISBN 978-953-51-0291-5
Hard cover, 138 pages
Publisher InTech
Published online 14, March, 2012
Published in print edition March, 2012

This book on Speech Processing consists of seven chapters written by eminent researchers from Italy, Canada, India, Tunisia, Finland and The Netherlands. The chapters covers important fields in speech processing such as speech enhancement, noise cancellation, multi resolution spectral analysis, voice conversion, speech recognition and emotion recognition from speech. The chapters contain both survey and original research materials in addition to applications. This book will be useful to graduate students, researchers and practicing engineers working in speech processing.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

N.R. Raajan, T.R. Sivaramakrishnan and Y. Venkatramani (2012). Mathematical Modeling of Speech Production and Its Application to Noise Cancellation, Speech Enhancement, Modeling and Recognition- Algorithms and Applications, Dr. S Ramakrishnan (Ed.), ISBN: 978-953-51-0291-5, InTech, Available from: http://www.intechopen.com/books/speech-enhancement-modeling-and-recognition-algorithms-and-applications/mathematical-modeling-of-speech-production-and-its-application-to-noise-cancellation

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen