INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

The Physiological Basis of Glottal Electromagnetic Micropower Sensors (GEMS) and Their Use in Defining an Excitation Function for the Human Vocal Tract

By

Gregory Clell Burnett

B.S. Physics (Southwest Missouri State University) 1991

M.A. Physics (Rice University) 1994

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Applied Science

in the

OFFICE OF GRADUATE STUDIES

at the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Committee in charge

January, 1999

Gregory Clell Burnett
March 1999
Department of Applied Science

The physiological basis of Glottal Electromagnetic Micropower Sensors (GEMS) and their use in defining an excitation function for the human vocal tract

Abstract

The definition, use, and physiological basis of Glottal Electromagnetic Micropower Sensors (GEMS) is presented. These sensors are a new type of low power (< 20 milliwatts radiated) microwave regime (900 MHz to 2.5 GHz) multi-purpose motion sensor developed at the Lawrence Livermore National Laboratory. The GEMS are sensitive to movement in an adjustable field of view (FOV) surrounding the antennae. In this thesis, the GEMS has been utilized for speech research, targeted to receive motion signals from the subglottal region of the trachea. The GEMS signal is analyzed to determine the physiological source of the signal, and this information is used to calculate the subglottal pressure, effectively an excitation function for the human vocal tract. For the first time, an excitation function may be calculated in near real time using a noninvasive procedure.

Several experiments and models are presented to demonstrate that the GEMS signal is representative of the motion of the subglottal posterior wall of the trachea as it vibrates in response to the pressure changes caused by the folds as they modulate the airflow supplied by the lungs. The vibrational properties of the tracheal wall are modeled using a lumped-element circuit model.

Taking the output of the vocal tract to be the audio pressure captured by a microphone and the input to be the subglottal pressure, the transfer function of the vocal tract (including the nasal cavities) can be approximated every 10-30 milliseconds using an autoregressive moving-average model. Unlike the currently utilized method of transfer function approximation, this new method only involves noninvasive GEMS measurements and digital signal processing and does not demand the difficult task of obtaining precise physical measurements of the tract and subsequent estimation of the transfer function using its cross-sectional area.

The ability to measure the physical motion of the trachea enables a significant number of potential applications, ranging from very accurate pitch detection to speech synthesis, speaker verification, and speech recognition.

Acknowledgements

This work was supported by the Lawrence Livermore National Laboratory, the National Science Foundation, and the UC Davis Department of Applied Science at Livermore.

This work would not have been possible without the efforts and enthusiasm of John Holzrichter, an associate director here at LLNL. He has worked tirelessly to promote this new sensor and its possible applications, and has been of invaluable assistance during the writing of this thesis. He has also suggested several experiments and has assisted with their implementation and analysis. He has done all this while continuing his work as a full-time AD and as a husband and father, and he is most appreciated.

I also owe a large debt to Larry Ng, my supervisor and group leader, who has spent much of his sought-after time helping me learn the intricacies of signal processing or just helping me solve problems, whatever they may be. Larry is a wonderful group leader, who knows when and how to lead, but also when to let things run their course. I couldn't have gotten through my four years here at the Lab without him and have been (and hope to be in the future) a proud member of his group.

I would also like to thank my friend and colleague Todd Gable, for many hours of useful (and not so useful, but always interesting) conversation and camaraderie. It would not have been possible to implement many of the experiments I conducted without Todd's assistance.

There are many people at LLNL that I would like to thank for their help: Greg Clark and Farid Dowla for invaluable assistance in the murky world of signal processing; Noel Sewall, for help with electronics and horses; Steve Patenaude, for being a good

friend, hangar buddy, and patient CFI; Brian Kolner, one of the smartest people I have ever met (no, he didn't see this before signing it!), for being on my committee and for many stimulating conversations; Jeff Kallman for his help with the 2-D E&M simulations; Rick Freeman, my thesis advisor and chair, for breathing new life into DAS and giving me plenty of good advice; Jong An, for arranging for me to use the Kodak EktaPro; and Roger Perry, who helped me set up and run the shaker experiment.

I would also like to thank my parents, Tommy Burnett and Jo Belle Hopper. They have always given me the love and support I have needed to get through this difficult portion of my life. My father instilled in me a love of working with my hands, and wouldn't let me just sit around and read when I was a kid. By taking the time to play with me and show me how to do things right the first time, he taught me the value of a well-rounded life. My mother was always there when I needed her and taught me the true meaning of compassion and sacrifice. She also bought me all the books I wanted and passed on to me her particular brand of witty sarcasm, to the everlasting chagrin of those around me. I couldn't ask for more loving or devoted parents.

My brother Jeff is the best brother a guy could hope for, and I miss our football games and camping trips together. He is a very talented athlete and a quiet man who enjoys helping others and protecting the innocent. He will be a great cop. My sister Jeni "Coach" Hopkins is a very successful athlete and coach, who has also excelled in her new job as Mama. I am very proud of both of them; they are both the best at what they do.

I would also like to thank the Cessna Corporation for the solid construction of N3781V, a 1949 140A in which I have spent many an hour getting the best therapy I

could ever buy. Also, I thank Bjorn Anderson for helping me keep 81V in the air and for teaching me to fly in clouds. A more patient and calm man the skies have never seen.

Finally, I would like to thank my partner for the last six years, Melinda Sue Bass. She is my first and only love, and has inspired me to do more and go farther. She is confident without being egotistical, and has a strength of mind and character few can equal. She is smarter than me in more than one area, but always makes me feel like I am the best. She will be one hell of a medical doctor, and I am proud of her intelligence, patience, and perseverance. I am honored that in six months she will take my name and I hope she will continue to put up with me for a very long time.

Table of Contents

List of Tables

List of Figures

Chapter 1

1.1 Foreword

I have striven to organize this thesis in a somewhat logical manner, but unlike a good novel the best cannot wait for the last – it has to be right up front, to make sure that you have done enough (for the professors tasked with reading it) to justify continued reading. Thus a large amount of suspense is lost, and keeping the reader's attention has been made that much harder.

The thesis begins (appropriately enough) in Chapter 1 with a justification of my work and its possible ramifications to society as a whole. That is, I suppose, what differentiates an applied physicist (as I view myself) from a conventional one – I prefer that my work have a timely and concrete influence on the world, rather than doing research for knowledge's sake. I think the latter is more noble and far-reaching, and that no significant advances may occur without it, but I do not have the patience for the value of the work to be recognized and I am far too goal-oriented to not be able to have something at the end of my labor that is not immediately useful. I am thankful we possess both types of researchers - as always, a good balance is essential.

In Chapter 2, I introduce the GEMS (and other radars) and explain how it detects motion. After the section on the GEMS, the tissues of the vocal tract are explored, and then the theory of sound production is detailed. A section on signal processing and transfer functions follows. Finally we discuss present audio-only techniques of approximating the transfer function and some current methods of imaging human tissues.

In Chapter 3, I discuss the experiments conducted to determine the validity of the three competing theories concerning the physiological basis of the GEMS return. The

first hypothesis (and perhaps most obvious upon initial examination) was that the return was due to microwave energy reflecting back from the folds themselves. A colleague postulated the second theory after the GEMS signal's similarities with an electroglottograph (EGG) were noted. This hypothesis has the return caused by transmission of a GHz wave through the folds, rendering the GEMS in effect another type of EGG. The third theory, put forward by myself, resulted from a more careful examination of the qualities of the GEMS return and of the geometry of the glottal area. This hypothesis postulates that the GEMS waves are reflecting from the front or rear wall of the trachea. Therefore, the return is due to the trachea ballooning in response to the pressure variations in the subglottal region caused by the modulation of airflow by the vocal folds. Thus the GEMS signal, while strongly correlated to fold motion, does not measure it directly.

In Chapter 4 I take the evidence from the previous chapter and use it to calculate an excitation function for the vocal tract. A lumped-element circuit model of the tracheal wall is constructed. The GEMS signal is then inverse filtered with the tracheal wall model to remove the tracheal wall effects and get back to the driving subglottal pressure. The subglottal pressure is then used as an excitation function of the vocal tract. This excitation function is then used with the recorded speech and a linear time-invariant model to obtain a transfer function of the vocal tract for voiced phonemes. A comparison of transfer function formant locations between individuals is presented which demonstrates how they may be used for speaker verification.

Chapter 5 contains a summary of all work completed and the inferences drawn from the results. I also have included suggestions for future work and some ideas for

utilization of this technology. Some of the projects have been worked on briefly in our lab for proof-of-concept purposes, but none as yet carried through to completion. It will take substantial efforts at universities and commercial research centers to complete them, but I am certain these ideas will migrate to the commercial community in the near future.

Completing the thesis are the Appendices. Here I have included two papers submitted for publication, plus a tutorial on inverting digital filters in the z plane that some may find useful. A list of phonemes of American English is also included, as well as a glossary, as readers will come from the speech, signal processing, and radar communities, each of which has their own vernacular.

1.2 Introduction

So why am I doing these experiments? Why am I taking almost four years of my life to study and answer these questions? Because I believe the outcome will significantly change our world for the better. The invention of this new type of micropower, radar-like sensors by Tom McEwan, the idea of the application of the sensor to speech by John Holzrichter and Larry Ng, and my work to understand the physiological basis of the return will affect our lives in many ways. One of the most significant will be the use of the GEMS to greatly improve the communication between man and machine.

Man's most rapid and efficient mode of communication is speech, but until recently it was unavailable for use with our machines. Even animals can respond to vocal commands, but we have had to make do with buttons, switches, keyboards, and mice to slowly interact with our machines. The first few halting steps taken in the speech recognition industry have given us an idea of the potential of man-to-machine communication but have also proved frustrating to use due to their inaccuracies, clumsy interfaces, and complete inability to work adequately in a noisy environment. That is about to change. Using this low power radar-like technology, I believe all of the present shortcomings of speech recognition and synthesis will be reduced to the point of nonexistence. We will be able to issue orders and receive spoken replies (even in a specific celebrity's voice) in noisy and unconstrained environments with ease. The entire relationship between man and machine is about to improve in a significant manner. This revolution is due to the ability of the GEMS to measure the effects of subglottal pressure on the surrounding tissues of the trachea. This measurement will, for the first time, allow

us to derive an excitation function of the vocal tract. Excitation function in hand, we can calculate the vocal tract transfer function, which enables many applications. We may now synthesize voices to a high degree of accuracy, examine the transfer functions to determine similarities (speaker independent recognition) and differences (speaker verification and identification), build phonetic recognizers (speech recognition), and many others. With the excitation function we may also determine the pitch easily and to a high degree of accuracy, as discussed in Appendix C.

This work is but the first of many on the subject that will be written. These low power sensors are so compact and useful that the number of possible applications will take many students and professionals alike many years to sort through, and indeed may be without end. However, the way in which the GEMS interacts with the human body must be firmly established, so that further research will have a solid foundation on which to build. There often exists a conflict between trying to make a new technology useful and understanding how it works. In this project, we were consistently fighting the pressure (both internal and external) to study and develop a particular application of the GEMS without really understanding the underlying principles of its interaction with the human body (and in particular the area of the neck surrounding the glottis). We were not always successful, but as a consequence my colleagues and I have been able to explore the potential of the GEMS in many different applications, which I will discuss at the end of Chapter 5.

However, the great majority of this thesis will concentrate on the physiological basis of the GEMS return – what exactly are we seeing when we observe the voltage return from the GEMS on an oscilloscope? Is it safe to use? Once we know what it is

and that it is safe, what useful things can we do with it? These are the questions I have set about to answer. I believe the answers hold great promise for mankind.

1.3 Overview of accomplishments

I have attempted to construct a sound foundation for the use of micropower electromagnetic sensors in speech technology. The acceptance of the use of GEMS and other EM sensors has been slow for many reasons: First, the ideas are quite recent, as John Holzrichter first envisioned the use of the GEMS for speech applications in 1994. Second, technological inertia, as many talented people have spent millions of man-hours on current audio-only technology and are reluctant to accept new technologies until they have been proven. There is also the dislike of the complication of adding another sensor to a communication package when the average consumer has trouble using a microphone correctly; fear of "microwaves"; and plain old stubbornness. I believe the early adopters of this technology will become the Intels and Microsofts of the speech communication world – I believe it will offer a great advantage in the near future. Like any other major shift in thinking, though, it will take time and a few vocal advocates in order for it to occur.

As for accomplishments that are a little more concrete, I have completed several. I have used several models (both physical and on paper) to show that microwaves at 915 MHz and 2.3 GHz reflect from the trachea much more easily than the vocal folds. I have designed and executed several experiments (along with my colleagues, without whom progress would have been significantly slowed or halted entirely) that have shown conclusively that the majority of the GEMS motion detected has been that of the tracheal wall and not the vocal folds themselves. I have constructed a lumped-element circuit model to determine the response of a tracheal wall to a contained pressure wave, and have used this knowledge to remove the dampening effects of the tracheal wall from the

GEMS signal to yield the driving subglottal pressure. This subglottal pressure may be used as the excitation function for a linear time-invariant model of the vocal tract, which in concert with the audio pressure recorded from a microphone may be used to derive the transfer function of the vocal tract. The transfer function is normally calculated every two glottal cycles (10-25 milliseconds), as the vocal tract does not change significantly over this time scale and the tract may therefore be approximated as invariant.

Even without using the audio signal, the GEMS gives us much more information than we have had available in the past – when the speaker starts and stops voicing, the pitch of their voiced speech (to a higher degree of precision and accuracy than with audio-only methods, see Appendix C), and information concerning the physiological vibration of the subglottal tracheal area. While the audio-only methods are able to do reasonably well in quiet environments (not as well as the GEMS can), their performance degrades considerably in the midst of merely moderate environmental noise. The GEMS works in the noisiest of environments without fail, as it is only concerned with the tissue motion of the person it is being used by, not the various noisy pressure waves surrounding it.

In order to place my work in the proper context and to facilitate understanding, I have included sections on the operation of the GEMS, an introduction into speech production and the vernacular of the speech community, and information on digital signal processing. I have also included some needed information on experimental apparatus.

Finally, I have listed several applications that could be supplemented with this GEMS technology. It is only a matter of gathering the capital and the people and these applications could be on every desktop in five years. I hope that we will find both and

get this technology out to the world. I believe it will make life more enjoyable and convenient, and isn't that the ultimate goal of all science?

-xxx-

Chapter 2. Introduction to the players

This chapter introduces the GEMS in the broader context of radar devices as a whole. It discusses how the GEMS detects motion of dielectric interfaces and how the signal from the GEMS is calibrated using an industrial "shaker". The tissues of the vocal tract are then presented, along with a discussion on the way humans produce sound. Finally, lumped-element circuit models, linear filter theory, and transfer functions are discussed in regards to the role they play in the modeling of the vocal tract and its walls.

2.1 The Glottal Electromagnetic Micropower Sensor

So what is this new, very low power, portable type of radar? Why is it referred to as a radar when it obviously has very little in common with the large rotating antennae and satellite dishes that we associate with RAdio Detection And Ranging?

The Electromagnetic Spectrum

The reason the GEMS is related to radars is that it transmits and receives waves in the "microwave" part of the electromagnetic spectrum. Figure 2.1 shows how the different parts of the spectrum are classified by wavelength. The microwave spectrum spans from roughly 3×10^{5} μm to 1×10^{3} μm, or 1 to 300 GHz. This is the same part of the spectrum in which "normal" radars operate. There are many further subdivisions within each part, but the gross structure is enough to understand at this point. The GEMS operates at about 10^{5} μm, on the low energy end of the microwave spectrum. The GEMS detects motion using (high frequency) radio waves, but does not yield range.

The microwave region is sandwiched between the meter-length radio band and the thermal infrared band. Microwaves can easily travel through smoke, fog, and haze that occlude the higher frequency infrared and visible light. Microwaves are used in cellular

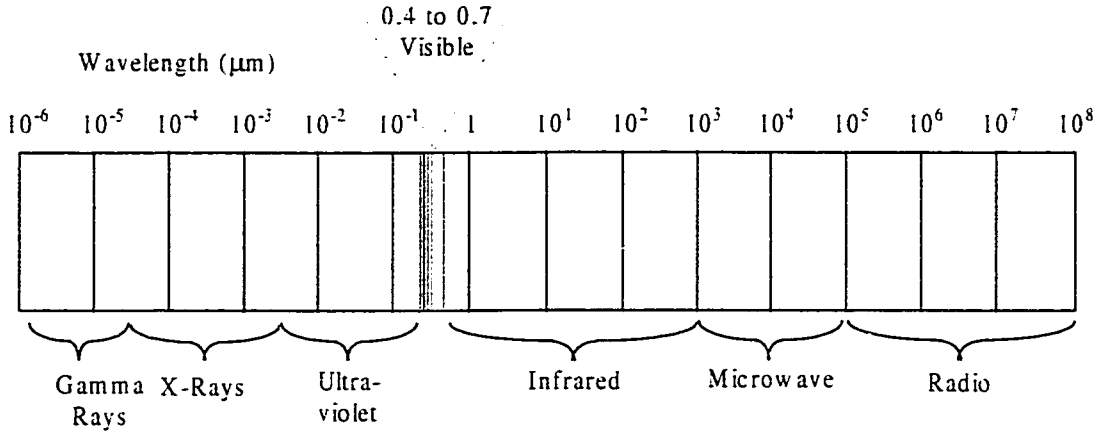


Figure 2.1. Classification of the spectrum based on wavelength

phones, telephones, television, and satellites because a smaller antenna (compared to one for radio waves) is needed for efficient reception and because microwaves can pass through the ionosphere, which will reflect most radio waves. It is the frequency range used by the Global Positioning Satellite system, which is so crucial and useful in airborne and nautical navigation.

2.1.1 Radar technology and the GEMS

Radar systems normally transmit pulses of energy several thousand or more times a second in the frequency bands spanning from about 1 to about 40 GHz. The same antenna normally acts as both transmitter and receiver, known as the monostatic configuration. These pulses usually consist of several cycles of the frequency in use, but can be as short as a single cycle. In the latter case they are referred to as either impulse radars or ultrawideband radars, as the uncertainty principle dictates that the shorter the pulse the more frequencies it must be composed of. Mathematically: $\Delta\omega\Delta t \geq 1$, where $\Delta\omega = 2\pi\Delta f$, the bandwidth, and Δt is the pulse width. Thus the shorter the transmitted signal the broader the range of frequencies transmitted.

Homodyne detection and field disturbance mode

In order to simplify processing, radars normally mix (multiply in time) the return signal with a second signal in order to translate the frequency content to a lower value to simplify processing. If the second signal is a delayed version of the transmitted signal, this is known as homodyne detection. If the second signal is generated separately, it is known as heterodyne detection. Both results allow the frequency content to be shifted because the multiplication in time is equivalent to an addition (or subtraction) in frequency. This will be discussed shortly.

Range is determined by the time it takes for the pulses to travel to the target and return. Angular location can be determined by rotating the radar beam physically or electronically through phased arrays. The accuracy of this measurement is determined by how well the radar beam can be focused. Speed is most accurately determined by the Doppler shift of the radar frequencies, or can be determined by the changes in position if the updates are rapid enough.

The GEMS is a motion detector that operates in a pulsed homodyne "field disturbance mode". In this method, anything that disturbs the reflectivity of the objects in the GEMS field of view causes a signal to be generated. For stability, filters are used so that only disturbances of sufficiently high frequency are registered – in our case, the highpass 3-dB frequency is about 80 Hz. Unfortunately, these filters cause distortion of the phase of the signal up to about 300 Hz. Since most male speech occurs in the 100-150 Hz range, this distortion must be removed through the use of inverse digital filters. The method used is discussed in Section 2.1.2.

The GEMS pulses are roughly 20 cycles long and are spaced 0.5 microseconds apart. Reflectivity information is derived using the homodyne method, resulting in a

signal that has frequency components at twice the frequency of the original and at DC. The amplitude of this signal is proportional to the phase difference between the transmitted and reflected pulse, which itself is a function of reflector distance from the antennae. This can be illustrated by the following example using sinusoids:

If we let the transmitted signal T(t) be a continuous cosine wave

$$T(t) = T\cos(\omega_0 t)$$

then the reflected signal R(t) will be

$$R(t) = R\cos(\omega_0 t + \phi)$$

where ϕ is due to the extra distance traveled by the transmitted wave and R and T are the amplitudes of the transmitted and reflected waves. The mixed signal X(t) will be

$$X(t) = T(t) \cdot R(t) = TR\cos(\omega_0 t)\cos(\omega_0 t + \phi)$$

or

$$X(t) = TR/2 \cdot [\cos(\omega_0 t - (\omega_0 t + \phi)) + \cos(\omega_0 t + (\omega_0 t + \phi))]$$

$$X(t) = TR/2 \cdot [\cos(\phi) + \cos(2\omega_0 t + \phi)]$$

where the first term is the DC component and the second term the component at $2\omega_0$. The DC component is usually selected out using lowpass filters. The result is a signal that varies in magnitude as $\cos(\phi)$ and is called the magnitude envelope M(t):

$$M(T) = TR/2 \cdot \cos(\phi)$$

The variation of the signal M(t) with phase ϕ (or distance) is graphed in Figure 2.2. A positive magnitude means that the presence of a reflector at that degree of phase difference will result in a positive voltage. A negative magnitude means just the opposite.

Finally, M(t) is filtered with a highpass filter (as described above) and amplified for output with a series of op-amps. Thus, only changes in the magnitude are amplified and output as signals. A DC signal indicates that nothing in the radar's field of view is changing or is in motion.

The highpass filter acts as a differentiator (i.e. a unity magnitude response and close to 90 degree phase response) for frequencies sufficiently above the 3-dB frequency. The phase response is 45 degrees at the 3-dB frequency, and how fast the phase increases to near 90 degrees is governed by the order and type of the filter. For the purposes of this demonstration, we will assume that for our frequencies of interest the highpass acts as a differentiator.

Therefore we differentiate the magnitude envelope to represent the amplification of only the changes in this voltage. It is important to note that this differentiation is an idealization of the filtering process described above. The differentiation of the magnitude envelope yields the sensitivity envelope, which is shown in Figure 2.3. It is plotted vs. free air distance to a reflector, assuming the index of refraction of the reflector is greater than that of the incident medium ($n_2 > n_1$). This assumption inverts the sensitivity envelope, as reflection from an interface where $n_2 > n_1$ causes the phase of the reflected wave to change by 180 degrees (see Section 3.2). If the reflection is occurring from a substance such that $n_2 < n_1$, the sensitivity envelope is not inverted.

The result of this is that the homodyne radar will have different sensitivities to motion depending on the distance away from the antennae. The same change in magnitude ΔM will yield different signals from the radar depending on how far the

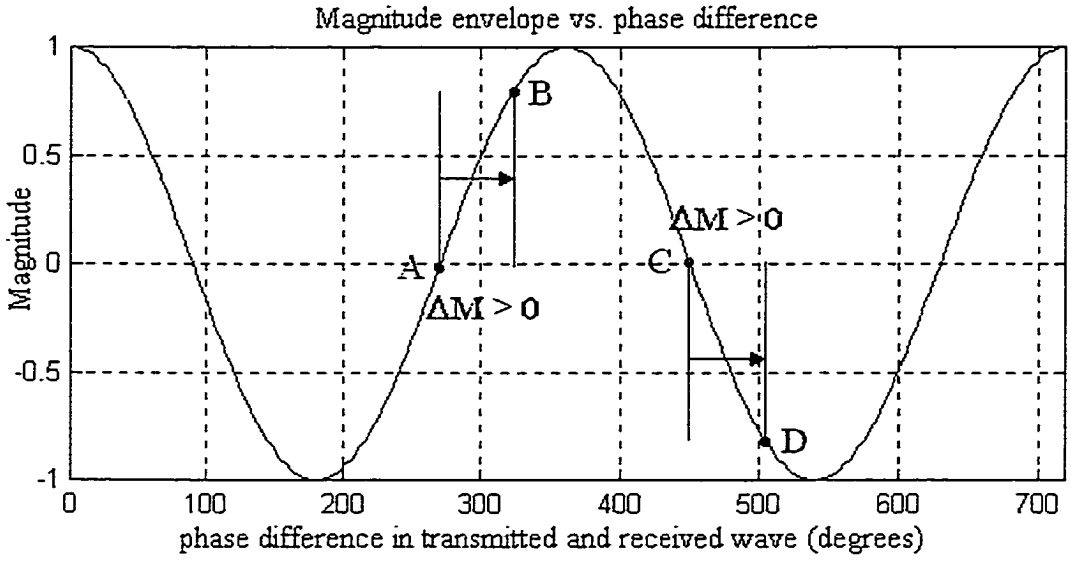


Figure 2.2. Theoretical magnitude response vs. phase difference ϕ in the transmitted and reflected waves for a homodyne system. The same change in position results in two different values for the change in magnitude depending on the phase difference between the transmitted and reflected waves (or distance to the reflecting interface).

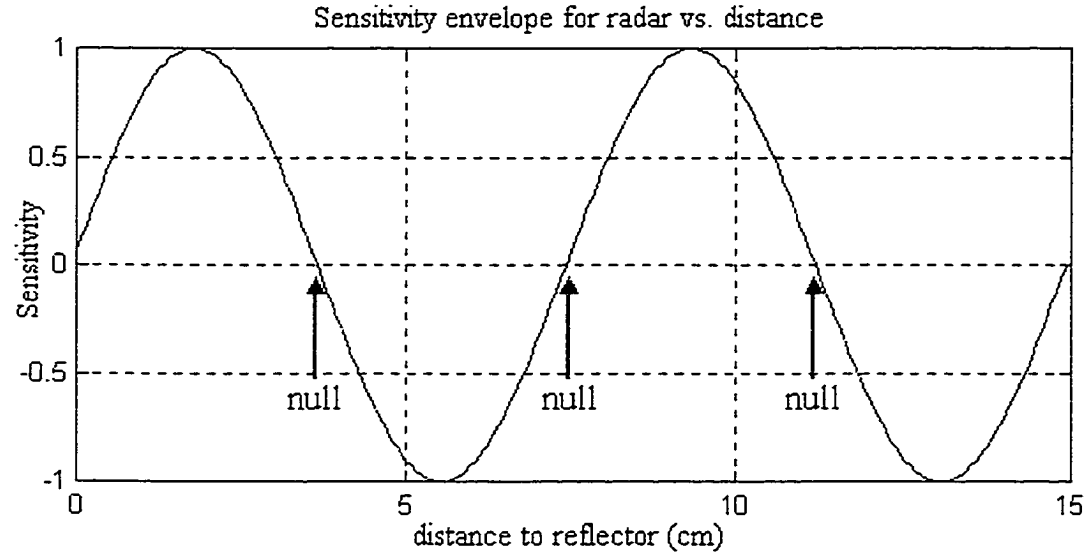


Figure 2.3. Theoretical sensitivity envelope for the radar as a function of distance away from the antennae assuming $n_2 > n_1$ reflection and a 15 cm wavelength.

reflecting interface is from the radar. This is illustrated in figure 2.2. Two changes in position are illustrated. It is assumed the change occurs very quickly, between subsequent transmitted pulses. The change in position from A to B and from C to D is the same, but C is a half-wavelength farther away from the antennae. It is clear that identical changes in position (denoted by the arrows) lead to very different changes in magnitude (in this example, one is positive and the other negative) depending upon the phase difference between the transmitted and received pulse (or equivalently, the distance to the reflecting interface). *Thus a reflector oscillating at a constant amplitude of motion will yield different radar signal amplitudes depending on its distance from the radar antennae.* The amplitudes will vary both in strength and sign. The frequency of oscillation is unchanged, but the amplitude of oscillation cannot be reliably determined. The GEMS can only determine amplitude of motion if the reflector is a known distance away from the antennae and the sensitivity envelope is known. Even then there are some distances ("nulls" located at $n \cdot \lambda/4$, in our example at 3.75, 7.5, 11.25, and 15 cm) where even large amplitude vibrations will result in a very small and/or distorted signal. They should therefore be avoided in normal operation. However, if their location with respect to a series of interfaces can be determined, the null positions are quite useful for establishing what the GEMS is reflecting from, as I will show in Section 3.5.

It should be noted that although the nulls will appear every quarter wavelength of the transmitted wave (15/4 = 3.75) away from the antennae, this does not mean they will be observed at the same distance away from the antennae. The distance will vary depending on the dielectric constant ε of the material through which the GEMS wave propagates. As will be discussed in Section 3.2, the waves are "compressed" by a factor

of $\sqrt{\varepsilon}$ in a dielectric material. Thus for an ε of 25, the nulls would occur at 0.75, 1.5, and 2.25 cm, which is the distances above divided by 5.

At the null distances, the response of the GEMS to even large velocities is very small, and if the amplitude of motion extends from one polarity to another, the return will be distorted. Thus to get an undistorted GEMS signal, the motion under study should be small in amplitude compared to the wavelength of the GEMS electromagnetic wave in the medium, and the range of motion should be confined completely to regions of either positive or negative sensitivity. For good signal strength it is preferable to be close to the points of maximum sensitivity. Also, for AC signals, the smaller the ratio of amplitude of the motion of the reflector to operational wavelength of the GEMS, the better the reproduction of motion. For small enough amplitudes, we are free to operate almost anywhere inside the sensitivity envelope. Since the wavelength of a 2.1 GHz signal is about 14 cm in air, if our amplitude of vibration in air is only a few mm, we are assured a linear signal if we are not close to a null. The vibration of the tracheal wall tissue is on the order of a millimeter (Sversky *et al.* 1997), so we may be confident of getting a good signal under most conditions.

One further point that has to be made clear: Unless the sensitivity envelope and the equivalent free-air distance is known, there is always an ambiguity in the sign of the GEMS vs. change in ϕ. As discussed above, in order to calculate the equivalent free-air distance in tissue, we multiply $\sqrt{\varepsilon}$ by the distance traveled in the tissue. This means that when the GEMS is reflecting off a structure in the body, we cannot be sure of the sign of the change in ϕ unless we know the thickness of the tissue and the index of refraction at about 2 GHz. It is an important ambiguity that must be kept in mind as we analyze the

data in Section 3.5. By obtaining a computed tomograpy (CT) scan of my glottal area, the absolute thickness of the tissue layers and therefore the equivalent free-air distance could be calculated, enabling accurate calculations of the sensitivity envelope inside the tissues of my body.

To assure a good GEMS signal, we could use a lower frequency EM signal, which has a longer wavelength. The longer wavelength would increase the distance between nulls, and we could then tune ϕ (the phase delay of the transmitted wave) so that the reflecting surface of interest would be far away from the null distances. This would ensure excellent reproduction of the motions under study. However, we are restricted in our choice of wavelength due to the effects of diffraction – once the cross-section of the objects we are trying to observe drops below a wavelength, the reflectivity of the objects drops markedly as the wave finds it easier to diffract around the object rather than scatter from it. This is illustrated in the electromagnetic simulations discussed in Section 3.2. We have conducted some GEMS experiments at 915 MHz, and they seem to perform well there. A 915 MHz wave has a wavelength of over 32 cm, resulting in a distance between sensitivity nulls of more than 80 mm in air (rather than about 35 mm for 2.1 GHz), which should be sufficient for excellent reproduction and much reduced positional sensitivity.

What does the GEMS signal represent?

Now that we have introduced the theory of operation of the GEMS, we can pose the question: What does the voltage coming from the GEMS actually represent? Well, assuming only a single reflector, the signal from the GEMS is a series of distance measurements (taken 2 million times a second) that has been highpass filtered. As such,

for frequencies below about 200 Hz, the GEMS signal resembles velocity. However, the effects of the highpass filters may be removed through noncausal digital filtering (explained in the next section), so we may get back to the original set of distance measurements. The digital model of the analog GEMS filters is a good fit for frequencies well below 50 Hz, so we may therefore conclude conservatively that for any frequency above 50 Hz to about 7 kHz (the measured upper 3-dB frequency of the GEMS's filters) we may recover the original position vs. time data. Thus the GEMS signal is a measure of position vs. time under the following conditions:

1) The motion is periodic with an amplitude $<< \lambda/4$, λ the transmitted wavelength
2) The center point of the motion is far enough away from the sensitivity nulls so that the amplitude of vibration does not extend over a null
3) The frequency of vibration is above 50 Hz and below 7 kHz
4) The frequency distortion due to the GEMS's filters is removed digitally

In all other cases the signal is not a reflection of the position, but rather some distorted version of it. For example, if the reflector is motionless, the output signal of the GEMS is zero volts, regardless of where it is located with respect to the sensitivity envelope. At this frequency (DC), then, the GEMS does not contain position information.

To conclude this section, I include Figures 2.4, 2.5, and 2.6. These illustrate the size and shape of the GEMS and its antennae, the measured cycles in a single pulse of the pulse train, the frequency spectrum of the transmitted GEMS pulses, and the antenna pattern measured for the simple rectangular antennae utilized by the GEMS that was used for the majority of this thesis work. The pulse train and frequency spectrum were

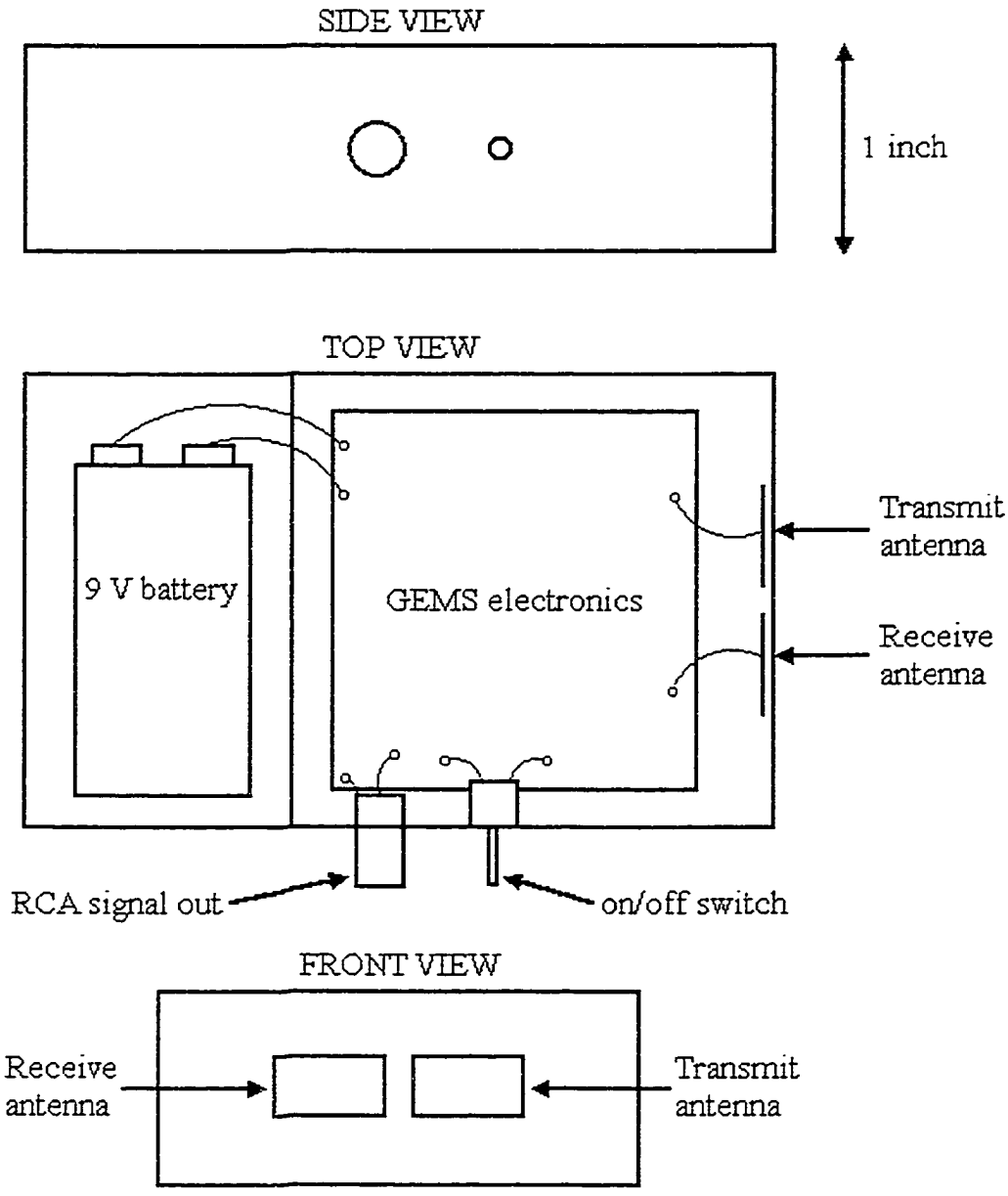


Figure 2.4. Overview of the GEMS's physical configuration.

measured with a Tektronix TDS 784A digital oscilloscope and a GHz band frequency spectrum analyzer.

The transmitted pulse has about 20 cycles, each about half a nanosecond long. This leads to a center frequency of about 2 GHz. The frequency spectrum has a maximum of about 2.1 GHz, with 3 dB frequencies at about 2.0 and 2.2 GHz.

The radiated antenna patterns were measured and the amplitude and power calculated by Bob Simpson and Doug Poland at LLNL. The patterns can be interpreted as follows: For the horizontal measurements the view is from the top, so 30 degrees means 30 degrees to the left of the GEMS in its normal operating position. For the vertical position, the view is from the right side, so 30 degrees denotes 30 degrees above the GEMS in its normal operating position. The units are in Volts.

The patterns show that EM energy travels in all directions, but preferentially directly above and behind the normal GEMS position. The horizontal pattern displays the GEMS's transmission asymmetry – the transmitting antenna is on the left side of the case (where the transmitting antenna is located, see Figure 2.4), and the receiving antenna is on the right side. Of most interest to us, though, is the pattern directly in front of the GEMS. It would seem that for our purposes the most energy directed toward the tracheal wall would be slightly to the left and down. There is more energy to the left and up, but it is at 30 degrees or more from the vertical while the energy lobe downward is quite close to the vertical, and so would have a better chance of geometrically scattering back to the antenna. However, it made no noticeable difference on the signal if the GEMS was held in its normal position or inverted.

From these measurements, it was determined that the peak power density transmitted from the GEMS at a distance of 3 cm from the antennae was 371 $\mu W/cm^2$, with a maximum time-averaged power of 7.4 $\mu W/cm^2$. Over all angles, the average peak power was 118 $\mu W/cm^2$ and the mean of the time-averaged power was about 2 $\mu W/cm^2$. The total peak power radiated into all quadrants was 13.4 mW, with an average total

power of 0.27 mW. These values are well below international continuous exposure limits, which we will discuss in Section 2.1.4.

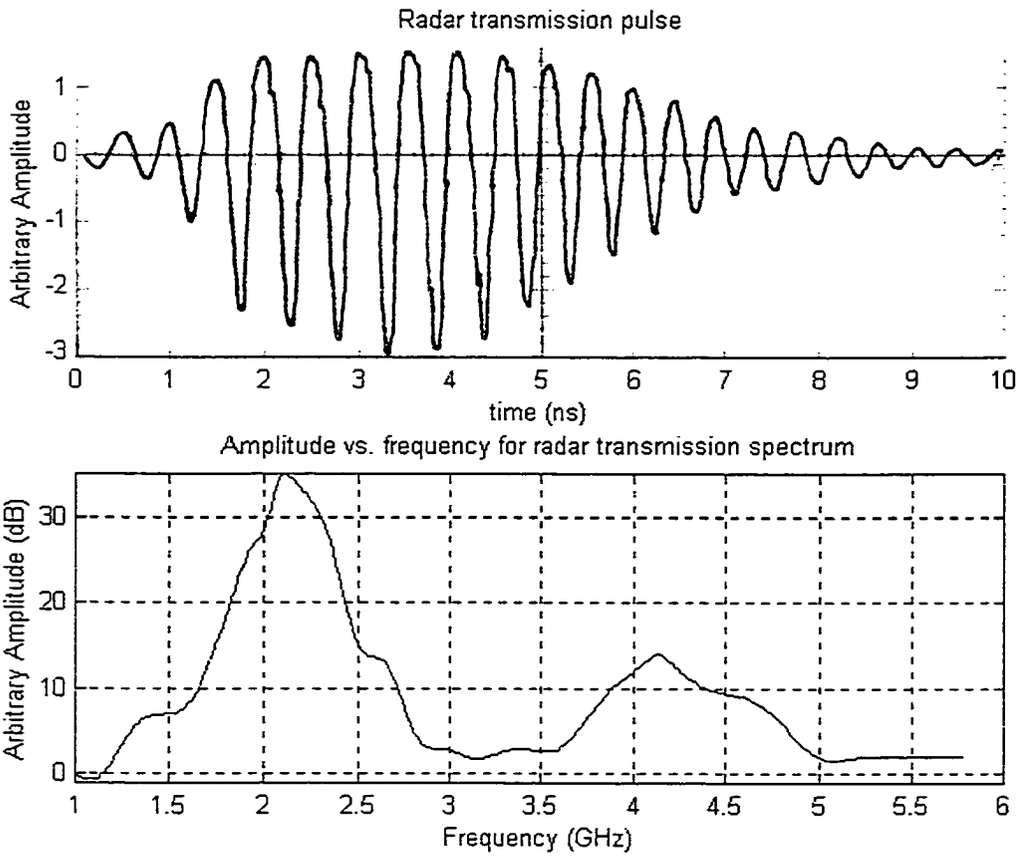


Figure 2.5. Plots of the GEMS pulse and frequency spectrum.

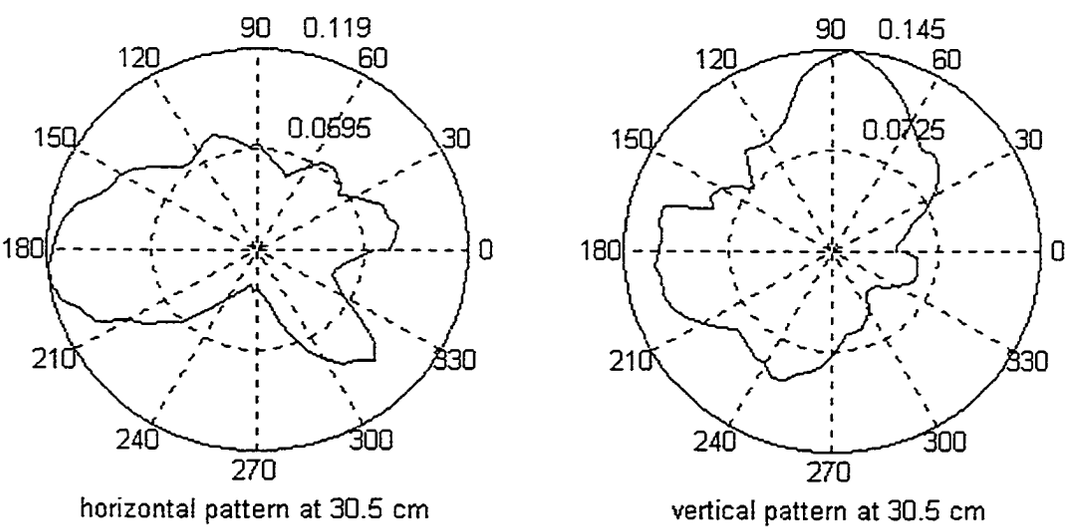


Figure 2.6. Measured antenna patterns for the GEMS antennae at 30.5 cm in Volts.

2.1.2. Removing the filter response from the GEMS signal

Statement of the problem

One of the first questions posed when we were first experimenting with the GEMS was this: Are we seeing the undistorted position return from the object that is causing most of the signal reflection or are the filters used in the GEMS distorting the signal significantly? We already know that the GEMS signal is representative of the position of a reflector, but is there anything else that could be affecting the signal? An even better question might be "Why are there any filters at all in the GEMS receiver?" We will answer the final question first and this might help us understand the first two.

The GEMS has filters in it so that we may have both a cleaner signal and higher resolution of the magnitude of the signal. The GEMS is limited to about ± 4 Volts in its output, and is quite sensitive to motion of any kind. The filters were necessary to get rid of all the low-frequency motion taxing place inside its "bubble" of sensitivity. The bubble for the GEMS in air is about 30 cm thick (starting from the surface of its case) and is distributed nonuniformly about the antennae (see the transmitted antenna pattern in Section 2.1.1). When the antennae are placed near the trachea, the GEMS is able to detect many kinds of low frequency motion: the jaw and tongue moving, the blood coursing through arteries and veins, the vertical motion of the glottis as speech begins and ends, and any motion of the GEMS itself with respect to the neck and trachea. Since we are interested in the vibration of tissues due to speech, the frequencies of interest are from 80 Hz up to about 5 kHz. Therefore a highpass analog filter was used with a 3 dB frequency of 80 Hz. Then the lower frequencies do not interfere with the signals of interest, and we may use the full ± 4 Volt range for the glottal signal.

In the future we are considering doing away with the analog filter and replacing it with a linear phase (no distortion) digital filter. However, this comes at a price in dynamic range, for if there are large low frequency signals the glottal signal will not have as much "room to expand". It is also possible that the low frequency signal would be large enough to saturate the GEMS's amplifiers, further distorting the signal. Figure 2.7

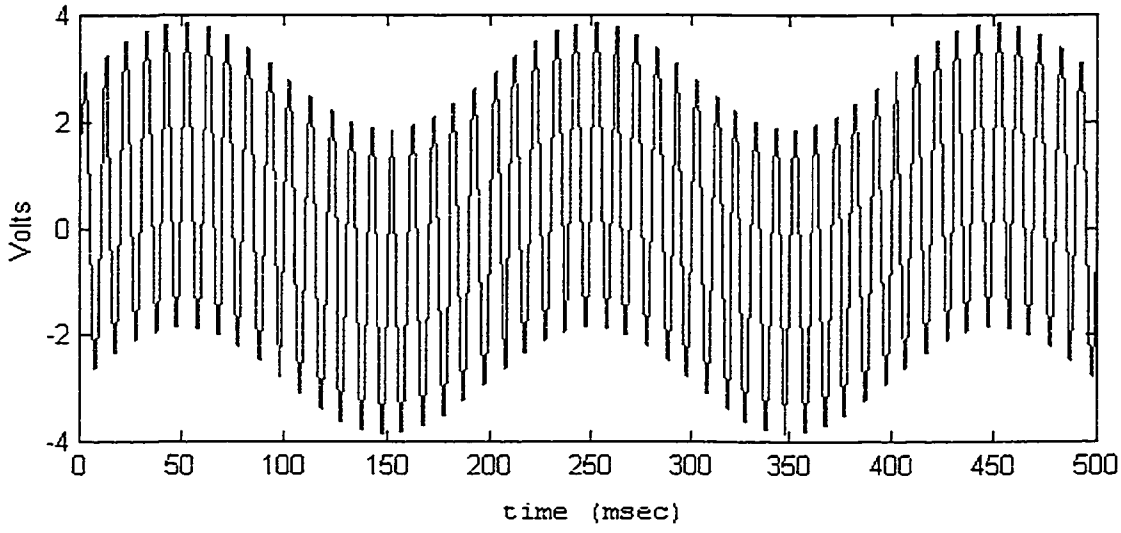


Figure 2.7. Example demonstrating how a low frequency signal can cause voltage amplitude resolution for a high frequency signal to degrade.

demonstrates how magnitude resolution may be lost. There is a 5 Hz signal with magnitude ± 1 V (perhaps the chin moving) and a 100 Hz speech signal "riding on top" of the 5 Hz signal at ± 3 V. Resolution can be lost if the overall signal amplitude requires a higher setting on the A/D. For example, if we use a 12-bit A/D, we have 4096 bins of magnitude resolution. This can be used for any of a number of ranges depending an the A/D and the application. Normal ranges are ± 5, ± 2.5, ± 1, ± 0.5, ± 0.25, and ± 0.1 Volts. For a ± 2 Volt glottal signal, a setting of ± 2.5 Volts on the A/D may be used and the resolution is 5 / 4096 = 1.22 mV/bin. For a ± 2 Volt glottal signal on top of a ± 1 Volt jaw signal, the next setting (± 5 Volts) on the A/D must be used. Then the resolution on

the signal is halved to 2.44 mV/bin. This loss of resolution would seem to discourage post-processing with digital filters. However, a digital filter has many advantages. A digital filter's parameters are easily manipulated and it can be designed to have linear phase, so there is no need for an anticausal post-filter to remove the distortion caused by the initial highpass filter. As long as reasonable magnitude resolution is available, it is a superior method. However, analog filters are so inexpensive and widely used they may continue to be utilized for some time.

So we now tackle the first question. Are the analog filters distorting the GEMS signal reflected from the tracheal wall? To get an idea of the amount of distortion we will determine the magnitude and phase response of the GEMS electronics. We can use this information to build a digital model that approximates the experimentally determined response. Finally, we digitally invert the model to construct an "inverse filter", which will be used to reconstruct the original undistorted GEMS position signal.

Determining the frequency response of the GEMS electronics

This is accomplished by injecting sine waves into the receiver end of the GEMS and then sampling the input and output. The magnitude and phase response at that frequency can then be calculated using a mathematical computational tool such as MathWorks' Matlab by measuring the frequency of the input and the resulting phase shift of the output. For accurate results, many different input frequencies should be used. In the experiment I conducted, a sweeping frequency generator was used to vary the frequency continuously between 20 Hz and 5 kHz. The total sweeping time was 12 seconds and was sampled at 20 kHz for good time resolution of the signals.

Good time resolution is needed to determine where the zero crossings occur in time so that the period (and thus the phase) for each cycle may be computed accurately. Digital systems are by necessity discrete, and this leads to errors in the period calculation on the order of ½ the sample period, in this case 25 μsec. This is not much at low frequencies, but can be substantial at high ones. For example, at 100 Hz (period of 10 msec) the relative error is about

$$e_r = \frac{.025}{10} = 0.25\%$$

But at 5 kHz the period is only 0.2 msec, or 4 samples. At this frequency the error can be as high as

$$e_r = \frac{.025}{0.2} = 12.5\%$$

with the error doubling as the sampling rate is halved. Fortunately, if you take enough points the response generally oscillates around an easily determined average and the response is not difficult to estimate. Linear interpolation of the zero crossing can also be used, but as the response changes smoothly the average was enough to get a good estimate of the filter qualities.

There are also errors introduced by the discretization of the signal amplitude, but this does not change with frequency and is not a large factor as long as the signal amplitudes use a good portion of the available resolution (for 12 bits that is ± 2048 samples). As an example, a signal that uses half the available resolution uses 2048 bins to represent the amplitude of the signal. The error in amplitude would be on the order of ½ a bin, so that the relative error would be only

$$e_r = \frac{.5}{2048} = 0.024\%$$

This is not large, and even to get 1% relative error would require using only 50 of the available 4096 bins, an unlikely occurrence. Thus most of the error inherent in this measurement comes from uncertainties in the time of occurrence rather than the amplitude of the signals.

The results from the experiment are shown in red in Figure 2.8. The trends for both magnitude and phase are quite distinct. They seem to indicate a simple one-pole highpass filter with a 3 dB frequency of perhaps 75 Hz. This fits in well with what we understand about the electronics of the GEMS – it should have a lowpass filter at about 80 Hz. Now for the challenge of building a digital model to emulate the performance of the analog filter.

Building the digital model

Modeling the continuous, analog world with a discrete, digital model is as much an art as a science. The best we can usually hope to achieve is a good approximation to most attributes of the analog world, usually glossing over the extremes in behavior and losing some of the subtleties.

There are many different types of LTI (linear time invariant) digital models, every one represented cryptically with an acronym. They include ARX, ARMAX, OE, and IV4 to name a few. These are all of the form

$$y[n] = \frac{B(z)}{A(z)} x[n - qk] + \frac{C(z)}{D(z)} e[n]$$

where x[n] is the input, y[n] is the output, and e[n] is noise in the system. The different models all are based upon the relations between the coefficients A, B, C, and D. For example, the OE (output error) model sets C and D = 1. The ARX model sets D = A and C = 1. There are many combinations possible and also some nonlinear models available,

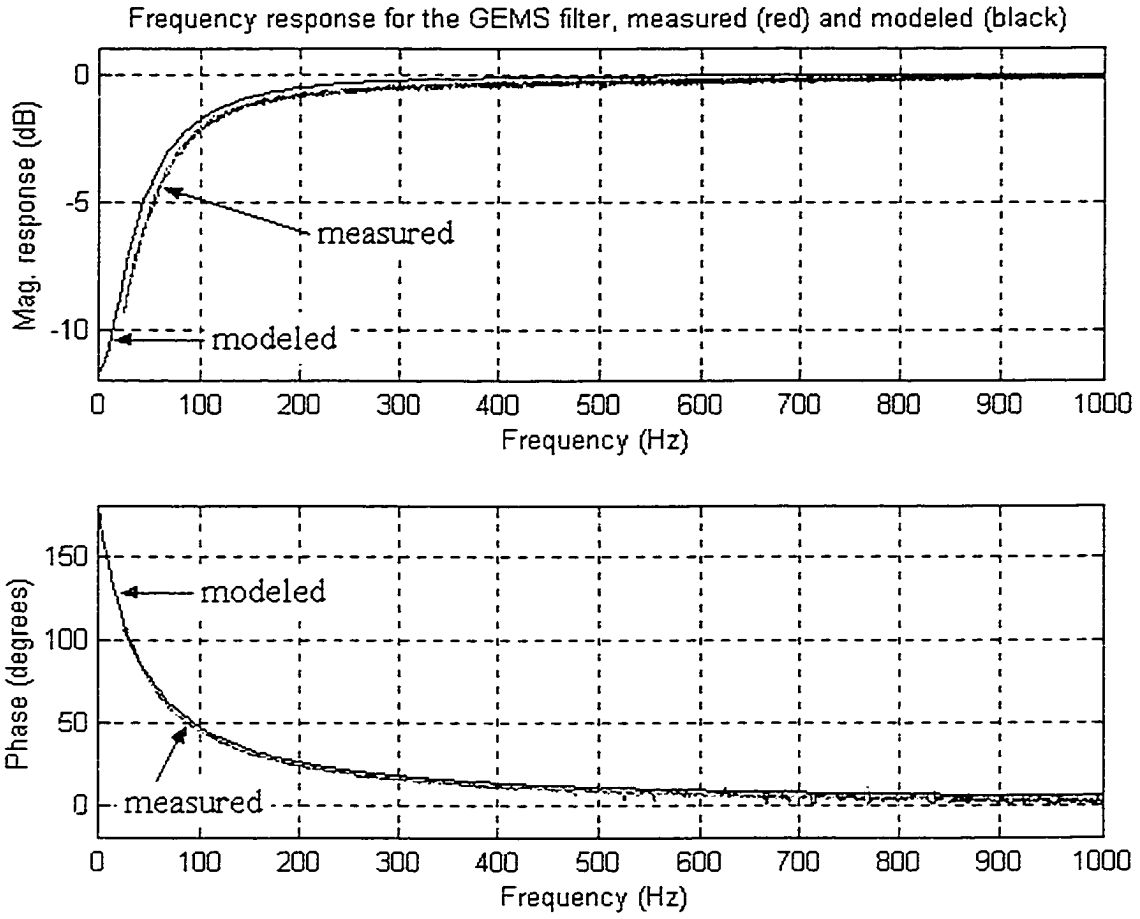


Figure 2.8. Measured GEMS filter response (red) and model response (black)

but as we are only trying to emulate a simple one-pole filter, a simple model seems to be in order.

Before I could build the simple model, though, I needed to know the approximate order of the model so as to limit my choices to a reasonable few. Using Van Den Boom et. al (1974) I was able to calculate the order of the A and B to about 2 and set C and D equal to 1. This in essence assumes the noise has no correlation with the input or output, which in our case is a good assumption as the plant is relatively simple and any noise in the system should be random. These orders indicate the output-error (OE) model would be the best choice. The OE model worked well with a single pole and zero (first order),

and with higher orders pole/zero pairs began to cancel, indicating we were using too many poles and zeros. Therefore the model used was a first order HP filter with a pole at $z = 0.9544$ and a zero at $z = 1.0123$. The model fit is shown in black in Figure 2.8. Note that the model fits the data quite well, indicating that only a simple filter is used in the GEMS.

Inversion and Stabilization of the digital model

The digital model response calculated above matches quite well with the measured data, but the model is not minimum phase – it has a zero at $z = 1.0123$, outside the unit circle. While this does not affect its frequency response, it does mean that it has no stable inverse. Thus the inverse filter must be stabilized using one or more allpass (AP) filters. The details of this procedure are given in Appendix A, so I'll present an abridged version here.

Inverting our model gives us an unstable filter with a pole at $z = 1.0123$. In order to stabilize the filter, an allpass (AP) filter with a zero at 1.0123 and a pole at 1/1.0123 is cascaded with the unstable filter. This cancels the offending pole at $z = 1.0123$ and yields a minimum phase inverse filter, by definition stable. In most applications, this would be the end. The magnitude is the same as the unstable filter (which is good) and we normally do not care about the phase. However, in our case the phase distortion is the main thing we are trying to remove so that we may associate physical tissue position with the GEMS return. Therefore we need to fix the phase.

The frequency response of the stabilized, inverted filter is shown in Figure 2.9. The magnitude is almost perfectly inverted but the phase is not. The phase is not negative enough, and at DC is zero degrees instead of –180. In addition, in Appendix A I will

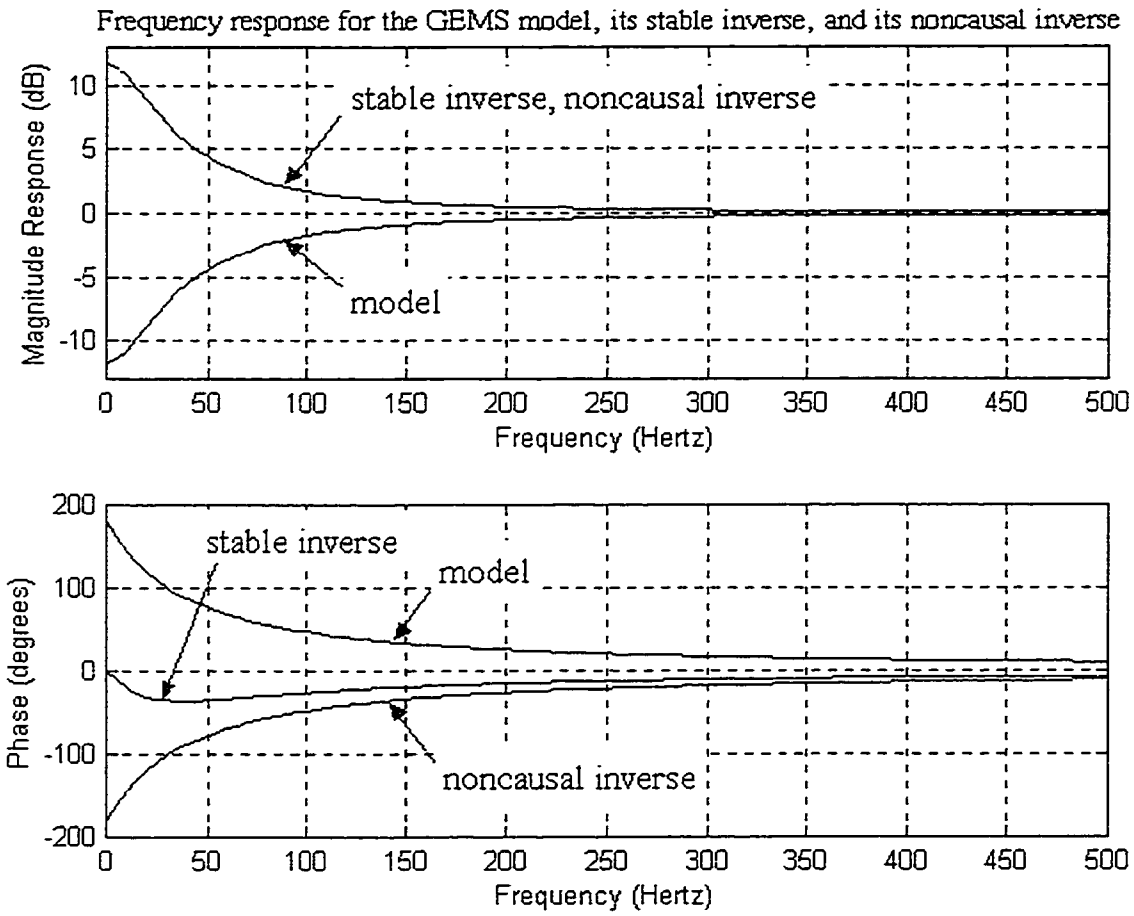


Figure 2.9. Magnitude and phase response for the GEMS, its stable inverse, and its noncausual inverse.

demonstrate how a stable allpass filter can never have negative phase at $\omega = 0$, so we can never build a filter that removes the phase distortion. Or can we? Clearly something else is needed.

Anticausality to the rescue

It is true that no causal filter can remove the phase distortion, and causal filters are what we are used to. Noncausal filters are given little coverage in most DSP and filtering classes, as they aren't "real" and cannot be implemented in real time. However, our processing of the GEMS signal need not be real time, especially at this experimental phase of development. Indeed, since we can cut the audio and GEMS signals precisely at

the beginning of the glottal cycle (the observed sharp drop in the GEMS signal, as will be discussed in Section 3.3), we may process each "window" of information after we receive it. Since the windows are on the order of 15-20 msec, there is not a large delay in the processing and transmission of the data. Thus noncausal filters are plausible (and in this case, the only) alternatives.

This "glottal synchronous" processing is a new method of audio processing made possible by the GEMS information. As the beginning of the glottal cycle is easily determined by the GEMS signal, all of our processing can be done in multiples of glottal cycles. As the data we are processing consists of entire cycles of waves, the fast Fourier transform (FFT) is more accurate and stable, resulting in more precise calculations. Conventional processing utilizes "sliding" windows, normally 30 msec in length. The length of the window is fixed, and it moves (slides) a fixed distance each processing step (normally 10 msec). The fixed window approach results in a heavy processing load and less accurate FFTs.

An anticausal filter is realized by taking the data to be filtered, reversing it so the first sample is now the last, and filtering it through a normal causal filter. The output is then reversed again in time to get the final result. The magnitude response of the causal filter is unchanged, but the phase response is inverted. To assure good phase distortion cancellation, then, we must find an anticausal allpass filter so that the combination of the causal, stable, inverse filter with the anticausal, allpass filter has the correct phase. For this simple single-order model, complete cancellation of the phase distortion should be possible.

Determining the characteristics of the anticausal all pass filter

Again, the details are handled in Appendix A, but basically it is as follows: A frequency is chosen where the match in phase is desired to be perfect. At this ω, the location of the pole and zero of the allpass filter may be determined using simple trigonometry. The results for the radar model are shown in Figure 2.9. It is clear that the noncausal combination of the stable, inverse filter and the anticausal allpass filter completely removes both the magnitude and the phase distortion of the GEMS's filters.

This success of this method only demonstrates that I have successfully inverted the original model, not that the original model is a good approximation of the filter or system. As I have shown in Figure 2.8 that the model is a good approximation of the GEMS's filters, and in Figure 2.9 that I can completely invert the response of the model, I feel confident that by using the noncausal inverse filtering process I am removing the effects of the GEMS's filters.

Therefore, we may feel secure that the inverse filtered signal is in fact very close to the true, undistorted position signal returning from the tracheal wall.

Differentiating to get velocity

With the analog filter distortion removed from the GEMS, we may now apply a differentiation to transform the position signal to a velocity signal. The velocity of the wall is necessary for the lumped-element circuit model used in Chapter 4. It seems simple enough – velocity is just the derivative of position:

$$v(t) = \frac{dx(t)}{dt}$$

However, building a digital filter that is a perfect differentiator is not so easy. The procedure is as follows:

The frequency response for the ideal analog differentiator is

$$H(\omega) = j\omega$$

The digital version, then, would be

$$H(\theta) = j\frac{\theta}{T}, \quad -\pi \le \theta \le \pi$$

where θ is the frequency variable in discrete time, and T is the sampling period (for those wishing a good explanation of digital signal processing, see Porat 1997). This can also be expressed in magnitude/phase form as

$$H(\theta) = \frac{\theta}{T} e^{j\frac{\pi}{2}}$$

In order to make the filter causal we include a delay of N/2 term (where N+1 is the number of taps) which manifests itself as a linear phase:

$$H(\theta) = \frac{\theta}{T} e^{j\left(\frac{\pi}{2} - \frac{N}{2}\right)}$$

We may use either a type III or type IV digital FIR filter. They are both antisymmetric (to get a difference), and type III has N even, with type IV N odd. To get the impulse response of $H(\theta)$, we apply the Fourier transform:

$$h[n] = \frac{j}{2\pi T} \int_{-\pi}^{\pi} \theta e^{j\left(\frac{\pi}{2} - \frac{N}{2}\right)} d\theta = \frac{(-1)^{(n-0.5N)}}{(n-0.5N)T} \quad N \text{ even}, n \neq \frac{N}{2}$$

If n = N/2, h[n] = 0 (Ibid.). We have only listed the response for even N (type III) as that is the only type where the phase distortion is easily removed. Type IV yields a slightly better magnitude response (Ibid.), but neither type has a flat 90 degree phase response due to the addition of the linear phase term. The linear phase may be removed for type III filters by cascading a second, anticausal filter in series with the differentiating FIR

filter. This anticasual filter simply advances the input by N/2, restoring the 90 degree phase shift. The result is shown in Figure 2.10.

Note there is some ripple in the magnitude response, but the phase response is exactly 90 degrees. Since we are more interested in the shape of the wave, and the ripple is less than 0.5 dB, having the exact phase response is more important than having the exact magnitude response. Thus we will use this noncausal filter when we want to differentiate the GEMS's inverse filtered position signal.

In conclusion, here are the steps for calculating the position and velocity of the reflector given the GEMS signal:

1) Take the GEMS signal of interest (it must be finite in length) and filter it with the stable, inverse filter.
2) Reverse the resulting data stream in time.
3) Filter the reversed stream with the allpass filter.
4) Reverse the data stream again. We now have an accurate position signal.
5) Differentiate the position signal using the noncausal differentiating filter above.

I have included a plot of the original GEMS signal, the inverse filtered GEMS (the position) signal, and the differentiated inverse filtered GEMS (the velocity) signal in Figure 2.11. It is clear that removing the filter distortion is important for getting the shape of the position curve correct, and the differentiation yields an interesting velocity wave – it would seem to indicate a more rapid transit to the negative position and a more leisurely climb to the positive position. We will investigate this velocity signal more thoroughly in the Chapter 4 when we relate it to the subglottal pressure.

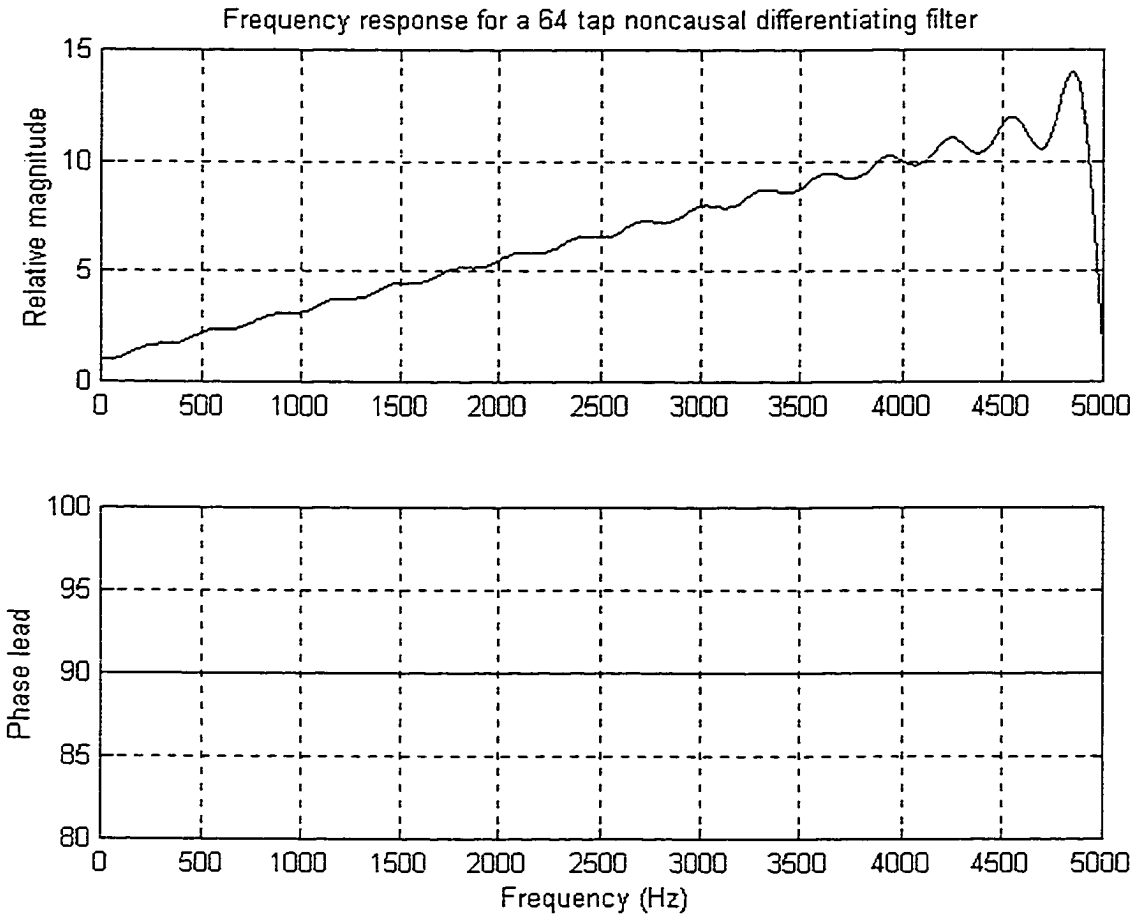


Figure 2.10. Magnitude and phase response for a 64 tap FIR noncausal differentiator.

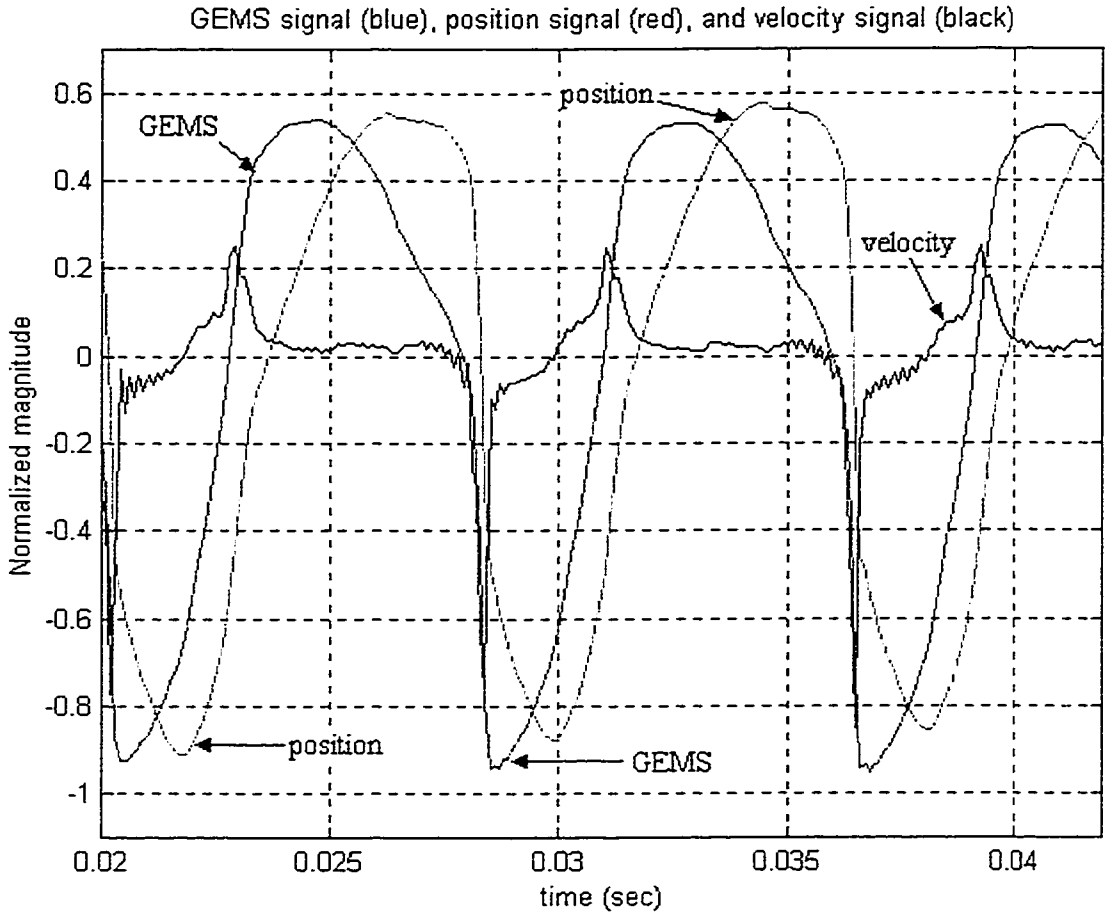


Figure 2.11. Plot of GEMS signal (blue), inverse filtered GEMS (position, red), and the derivative of the position (velocity, black) for a normal chest phonation.

2.1.3. The shaker experiment

As the GEMS unit designed for our use in glottal experimentation was not designed specifically for physiological applications, we conducted several experiments to determine its operating characteristics – how it responds to known vibrations at different distances and amplitudes. For this we turned to an industrial-strength calibrated "shaker" – essentially a piston-like mechanical linear oscillator about 35 cm in diameter driven by magnetic coils. The frequency and amplitude of vibration can be easily controlled, and the frequency can be swept from one frequency to another in controllable steps over several seconds. Amplitude is measured and controlled using an accelerometer. From the data gathered and knowledge of the phase and magnitude characteristics of the equipment, we were able to determine to a large extent how the GEMS units detect motion.

Description of experimental setup

The shaker itself was a large, cylindrical shaped device with a solid metal top (similar to a drum) overlaid with a thin layer of rubber. This "drumhead" was approximately 35 cm in diameter. A solid aluminum block with a surface about 10 cm square was mounted to this drumhead with 4 bolts. An accelerometer was mounted on the upper right corner. It was cylindrical with a length of 3 cm and a diameter of 1.5 cm. The GEMS was suspended from a system of adjustable rails and was attached to a platform that could be raised and lowered with a hand crank, ensuring accurate placement. The rails were placed on top of some foam in order to damp the vibrations of the floor induced by the shaker. As a result the GEMS was suspended vertically above

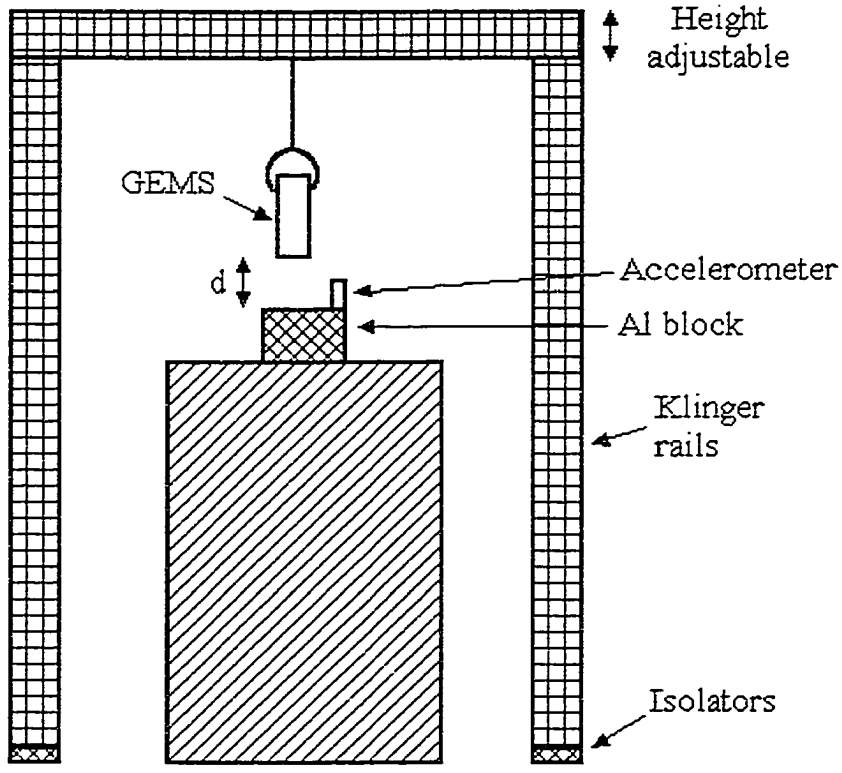


Figure 2.12. Shaker experimental setup.

the aluminum block in a vibration-free position. Care had to be taken to protect the data cable coming from the GEMS from vibrations as well. See Figure 2.12.

The accelerometer was a Dytran model 3134A, serial # 243. It had been calibrated several times in the last 10 years and its frequency response was well known. Between 10 and 1000 Hz, the magnitude and phase response were essentially flat. Thus we can be assured that the position data derived from the accelerometer (through integration, see below) was accurate.

The GEMS and accelerometer signal were simultaneously recorded using a PC laptop with a Labview 4.0. A National Instruments DAQCARD AI-16E-4 A/D with a shielded BNC-2080 connector board was used to connect to the accelerometer signal (channel 0) and GEMS output (channel 1). The shaker frequency used for all

experiments was 100 Hz and the sampling frequency was 40 kHz with no prefiltering. The high sampling rate assured that no aliasing would occur. For processing, the data was read into Matlab 5.21 and lowpass filtered to 4.9 kHz using a high order distortion-free digital Chebyshev-II filter. It was then decimated by 2 to 20 kHz and then again to 10 kHz. The GEMS signal was inverse filtered to remove the distortion at 100 Hz caused by the GEMS's internal filters.

The shaker was driven at 7.9 g's using a driving voltage about 3 V_p. Displacement was initially estimated via a calibrated table to be ± 0.008 inches, or ± 0.20 mm, and was later confirmed by integrating the accelerometer signal twice and setting the integration constants to zero:

$$\text{distance} = \frac{-a \cdot \left(9.8 \times 10^2 \text{ cm/sec}^2\right)}{4\pi^2 f^2} = \pm 0.20 \text{ mm}$$

where a is the acceleration in g force (7.9) and f is the driving frequency (100 Hz).

Distance Sweeps

In these experiments the distance from the GEMS's plastic case to the shaker head was measured with the shaker deactivated. We began with the distance at 15 mm for the first run and increased by 5 mm intervals up to 305 mm. The GEMS signal was observed to vary in both amplitude and composition, occasionally exhibiting slight distortions, probably due to the saturation of the GEMS's amplifiers. These distortions were observed at distances of 15 mm and between 65-95 mm. In addition, extra recordings were made at the distortion points at different amplitudes to see if the amplitude of the motion affected the distortion. In total 4 distance sweeps were recorded on 4 different days.

In order to compensate for this diminished subtended area by the block as the distance between it and the GEMS is increased, I first estimated the distance at which the GEMS pattern filled the entire shaker head. The first results indicated a sharp drop-off in magnitude response at a distance of about 100 mm. A little trial and error fixed this baseline distance at 85 mm. At this distance it as assumed that all the energy released in the forward direction by the GEMS is reflected back to it by the shaker head. Upon reflection, the energy returns to the GEMS where some of it is captured by the receiving antenna. The amount received is proportional to $1/r^2$, where r is the round-trip distance. When r is increased from this baseline distance, the energy is not completely reflected back to the antenna, as the shaker head no longer fills the field of view of the GEMS. To calculate the new area seen by the GEMS, we need to know the angle between the GEMS and edges of the shaker head at the baseline distance:

$$\Theta = \operatorname{atan}\left(47/85\right) = 28.9 \text{ degrees} = 0.531 \text{ rad}$$

where 47 mm is half the width of the shaker head and 85 mm is the baseline distance used.

The next step is to calculate the area "seen" by the GEMS at the new distance:

$$A(r) = \left(2 \cdot r \cdot \tan(\Theta)\right)^2$$

We then take the ratio of the area seen at the new distance to the actual area of the block, in this case 8035 mm^2, to get the correction factor:

$$C(r) = \frac{A(r)}{8035}$$

Using this correction factor, the inverse filtered GEMS, and the position signal derived from the accelerometer, I calculated the sensitivity curve shown in Figure 2.13:

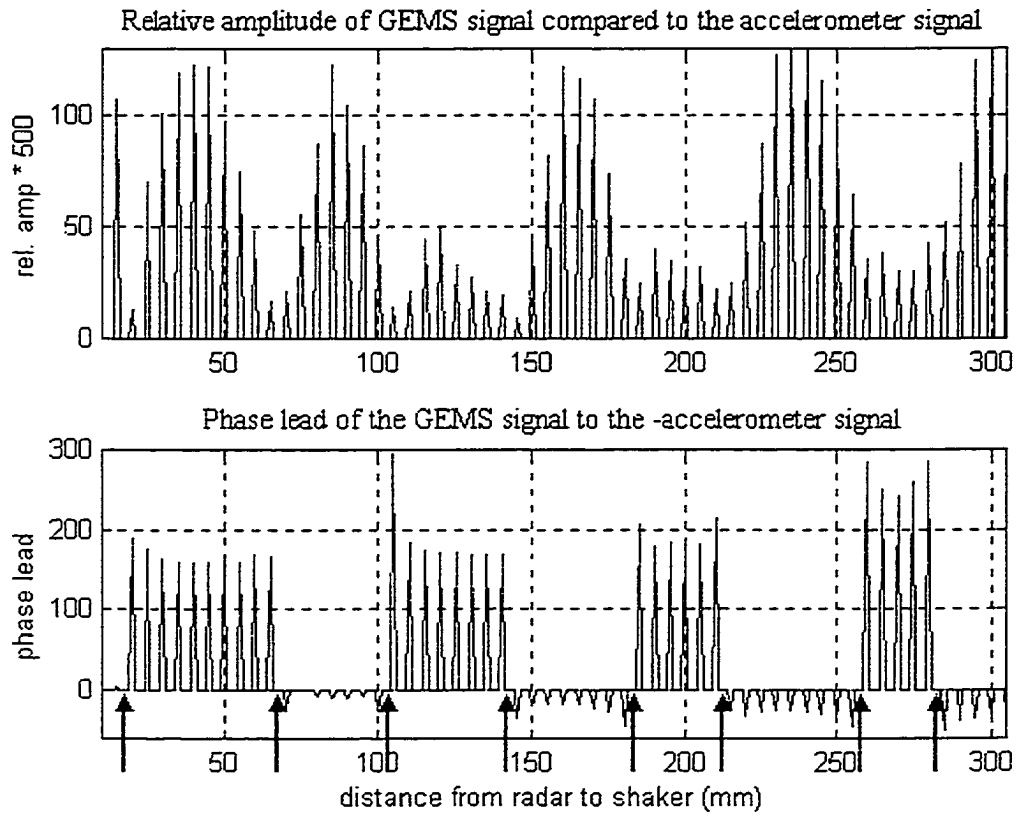


Figure 2.13. Relative amplitude (GEMS/accel) and phase lead of the GEMS vs. distance from the GEMS to the shaker block. The arrows denote points where the GEMS changed phase with respect to the inverted accelerometer signal.

It is clear that the GEMS changed sign with respect to the accelerometer position signal, signifying a change in phase between the GEMS and the accelerometer of 180 degrees, about every 35-40 mm, or $\lambda/4$. The more irregular phase shifts observed at 105 mm and longer distances are the result of very small and distorted GEMS signals, which make the phase difficult to determine accurately. The phase is calculated by comparing zero crossings, so any noise in the signal (or just a very weak signal) degrades the accuracy of the calculation. The distance at which the GEMS changes sign corresponds to the sensitivity null distances, discussed in Section 2.1.1. The nulls are located $\lambda/2$ apart, but since a reflection occurs the distance the wave travels from transmitter to receiver is twice the distance to the reflector. Thus the one-way distance between nulls is

λ/4, which for 3.5-4 cm corresponds to a frequency of about 1.9-2.1 GHz. This agrees well with the 2.1 GHz center frequency determined by the spectrum analyzer in Section 2.1.1.

The accelerometer was calibrated so that a motion upward (toward the GEMS) was defined to be positive, and the GEMS's phase was interpreted in that manner. The results are shown in Figure 2.14. This is the sensitivity envelope from the previous section with the phase information integrated into the envelope. A negative sensitivity means the GEMS's polarity is opposite that of the motion. For example, at a distance of 85 mm in

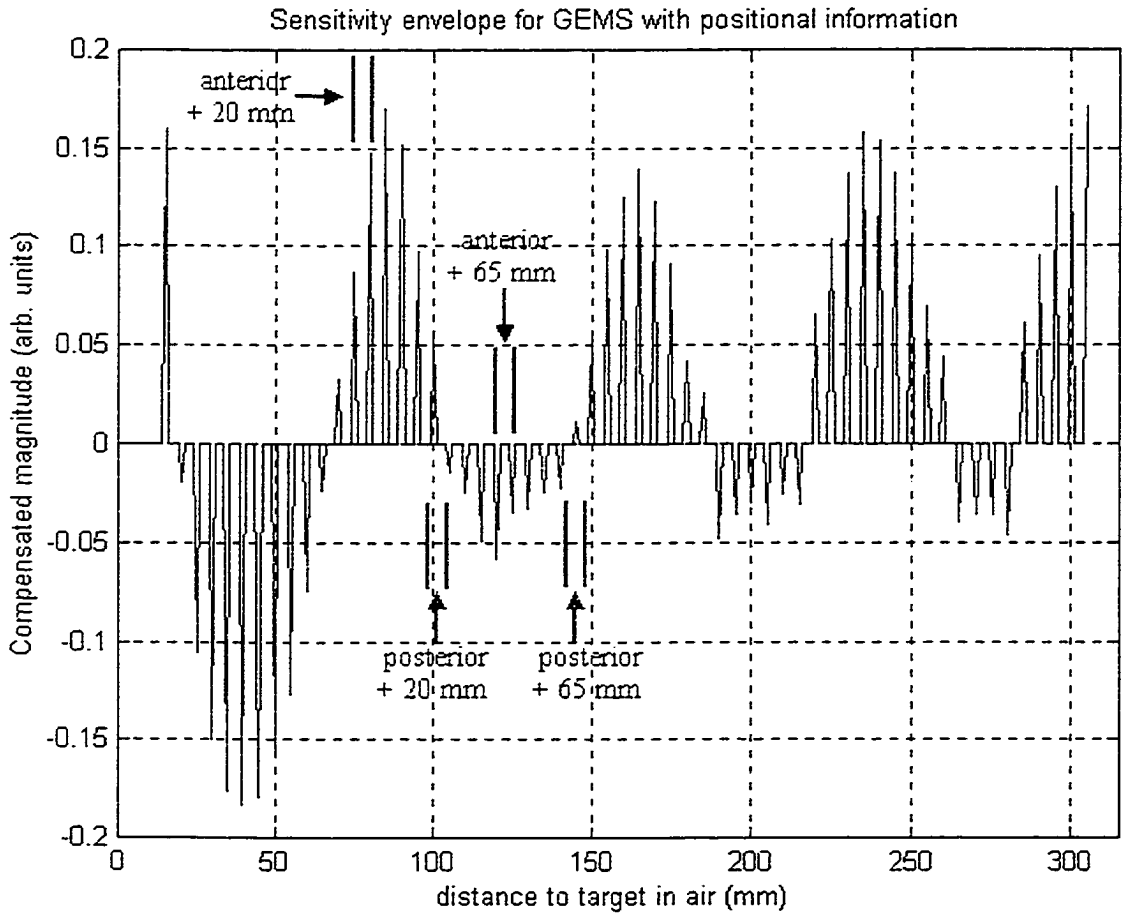


Figure 2.14. Sensitivity envelope of GEMS. A positive sensitivity denotes a positive signal for a reflecting surface moving toward the GEMS. The calculated positions of the null points for the anterior and posterior wall are shown in blue. They are discussed in Section 3.5.1.

air from the antennae, a positive GEMS signal would correspond to a movement of the reflector toward the GEMS. At 50 mm, the same motion toward the GEMS would result in a negative signal of about the same amplitude. In both cases this assumes the second material has a larger index of refraction than the first, i.e. $n_2 > n_1$. If this is not the case, the motions will be in the opposite directions due to the lack of a 180 degree phase change upon reflection. Thus the use of the accelerometer-derived position gives us access to absolute motion directions for the first time, and we will use this knowledge in Section 3.5 to determine whether the anterior or posterior wall of the trachea is causing the majority of the signal return.

GEMS signal strength vs. amplitude of vibration

It is instructive to compare the GEMS signal strength noted at a distance of 85 mm from the shaker to the signal observed from my glottal area. At the sensitivity maxima located at 85 mm, the GEMS signal was about 1.3 V peak. The displacement (calculated from the accelerometer) was about 0.4 mm peak-to-peak. The area of the reflector is not important as long as it is much larger than the size of the antennae, which are each only about 1.5 cm by 0.8 cm. This is because when the reflector re-radiates the reflected energy it acts as a transmitter. The GEMS receives that energy on its receive antenna. As long as it is much larger than the receive antenna, the size of the reflector is irrelevant as long as the reflected wave is large enough in extent to fill the receive antenna. Thus for comparison of signal strengths at the same effective distance away from the GEMS, we may ignore the total area of the reflector and focus on the linear vibrational amplitude.

The GEMS, at 85 mm, assuming 100% reflection, returns a signal strength S of

$$S = \frac{1.3\,V_p}{0.4\,mm} = 3.25\frac{V_p}{mm}.$$

When I take GEMS measurements of my own rear tracheal wall (which happens to be located very near to 85 mm in free air, see Section 3.5), I get about a 4 V_p signal. If we use S above, neglect conduction losses, and assume the reflection is only about 32% efficient in amplitude (it has to endure two 43% efficient transmission through a muscle-air interface and one 57% efficient reflection from a air-muscle interface, resulting in a reduction to about 11% in power, or 32% in amplitude - see Section 3.2 for information regarding reflections at interfaces), this indicates an amplitude of vibration of about

$$\text{Amplitude} = 0.32 \cdot \left(\frac{4\,V_p}{3.25\,V_p/\text{mm}} \right) = 0.4\,\text{mm peak},$$

in good agreement with the estimated 1 mm peak amplitude of the vocal tract calculated by Svirsky et al. and certainly within the realm of possibility.

In conclusion, the distance sweep showed us that the strength and the sign of the GEMS return does vary approximately sinusoidally with a wavelength of about 7.5 cm. After compensating for the effects of moving a finite-sized target away from the source, the relative amplitude for positive sensitivity was relatively constant. Also, the null distance locations in free air were established, which we will also use in Section 3.5 to determine where in the neck the reflections are taking place. At maximum sensitivity, the GEMS signal strength is about 3.25 V_p/mm.

Minimum Detectable Amplitude

The first maxima of the GEMS sensitivity envelope, at about 40 mm, was chosen as the distance at which to determine the minimum amplitude of motion detectable by the present incarnation of the GEMS. The smallest amplitude detectable by eye on an oscilloscope (which had a signal to noise ratio on a power spectrum of approximately 16

dB) was at ± 4.2 μm. At this displacement the GEMS signal was about 12 mV, and the accelerometer signal was about 66 mV, leading to a position amplitude of ± 4.2 μm. Since the sensitivity at 40 mm is close to that at 85 mm, we may use the sensitivity of 3.25 V_p/mm above to calculate the peak amplitude according to the GEMS:

$$\text{Amplitude} = \frac{1.2\times10^{-2}\ V_p}{3.25\,V_p/\text{mm}} = \pm 3.7\,\mu\text{m},$$

in good agreement with the amplitude calculated from the accelerometer data. Thus the GEMS is capable of detecting motions on the order of micrometers, depending on the reflectivity of the object, its frequency of vibration, and its distance away from the GEMS. This is truly remarkable for such a low-power, portable device.

2.1.4. Safety issues

Is it safe?

This is one of the most often asked questions I get when I demonstrate the GEMS. In short, absolutely. The GEMS stands for Glottal Electromagnetic Micropower Sensor, and it is just that – micropower. It emits pulses of microwave energy so low they cannot be measured by normal radiation detection equipment. The Hazards Control Department at LLNL was called in to do the initial safety measurements in 1996 as part of the protocol for beginning studies with human subjects. They found that for the GEMS the emissions were "too low to measure" using the most sensitive equipment available to them. Recently (see Section 2.1.1) measurements of the radiated power were completed here at LLNL using very sensitive equipment. At a range of 3 cm, they found the following:

Center frequency = f_c	= 2.1 GHz	
Pulse duration = t_p	= 10×10^{-9} seconds (how long the pulse lasts)	
Repetition rate = r	= 2×10^{6} Hz (how often the pulse is repeated)	
Duty factor = d	= 0.02 (how much of the time energy is emitted, $d=t_p r$)	
Max ave. power measured P_a	= 0.007	mW/cm^2 (maximum)
Max peak power $P_p = P_a / d$	= 0.35	mW/cm^2 (maximum)
Peak energy $E_p = P_p \times t_p$	= 3.5×10^{-6}	$\mu J/cm^2$ (microjoules per square cm)
Electric field = $(3770 * P_a)^{1/2}$	= 0.016	V/m (Volts/meter)
Magnetic field = $(P_a/37.7)^{1/2}$	= 0.043	mA/m (milliamps/meter)

These are all calculated using the maximum value measured from the GEMS. We can then compare these emission levels to several different standards including the most stringent in the Western world (Finland's STUK) and the IEEE/ANSI American standards as well as LLNL's own standards for non-ionizing radiation. The results are paraphrased in Table 2.1. It is evident that the exposure due to the maximum estimated power is a small fraction of all of these standards.

Type of exposure	Strictest standards*	IRPA/LLNL standards**	GEMS emissions
Electric fields (V/m)	19 (Sweden)	61 (IRPA)	0.162 (calculated)
Magnetic fields (A/m)	0.05 (Sweden)	0.16 (IRPA)	0.00043 (calc.)
Power density (mW/cm^2)	0.1 (Sweden)	1.00 (IRPA)	0.007 (measured)
Peak Energy ($\mu J/cm^2$)	0.6 (Finland)	3.3 (LLNL)	0.0000035 (calc.)

Table 2.1 Continuous exposure limits for 2.5 GHz electromagnetic radiation for the general public.

* Standards are taken from North America and Western Europe, supplied by G. Miller, LLNL Hazards Control Department

** IRPA is the International Radiation Protection Association. LLNL is the standard used at the lab.

As you can see, the GEMS is quite acceptable by all standards for continuous exposure for microwaves. But, in this era of eyeing the government suspiciously that does not seem to be enough. Is it really safe?

Well, let's compare it to something we're familiar with. Everyone knows that microwave ovens put out microwave energy – how much do they expose us to? Well, according to some Siemens researchers trying to detect how much electromagnetic interference comes from common household devices (Faber & Rybinski), the following values were obtained:

Measured Field Strength	(Volts/m)
Electric Hand Drill	1-2
Transceiver Set	3-18
Fluorescent Light	1-3
Microwave Oven	1-3

Table 2.2 Measured electric field intensities for common devices (Faber & Rybinski).

The GEMS's maximum calculated value of 0.016 V/m is much lower than any of these common items. Another valid comparison is that of cellular phones. They operate in a similar part of the spectrum and transmit microwaves at a power of about one watt, about one hundred times the maximum power radiated by the MIR. So we might more cautiously conjecture that the GEMS is as safe as any of these devices and is likely to be much more so. If you can tolerate fluorescent lights and cell phones, you should be unaffected by the GEMS.

Chapter 2.1.5. Previous tissue work using microwaves

There have been some attempts in the past to use microwaves to image, find the transmittivity, or sense movement of portions of the human or animal body. Nothing quite like we have been doing with the GEMS, but nonetheless it may be instructive to review the literature.

Much of the past work has been summarized in the recent book "Medical applications of microwave imaging" (1986, New York: IEEE Press, ISBN: 0-87942-196-7). There has been an upturn in interest in the use of microwaves to probe the human body in the last 15 years, due to the emergence of fast (picosecond) circuitry (that has made microwave transceivers smaller and more affordable) and the advent of small, powerful computers. A search of the literature revealed few transceivers similar to McEwan's invention and no attempts at all to try and detect the motion of any speech articulators, much less attempt to measure the vibration of the subglottal region and associate it with pressure below the glottis.

However, there are papers devoted to microwave radiometry (tissue temperature measurement; Meaney et al. 1996; Mizushina et al. 1995), detection of cardiac and breathing motion (Chan and Lin, 1987), the imaging of tissue using both transmission through and reflection from the body (Skolnik, 1986; Young and Peters, 1986; Larsen and Jacobi, 1986; Jacobi and Larsen, 1986; Lin.1986; and Murphy 1994), and general information and encouragement (Burdette et al. 1986; Land 1995; Boerner and Chan, 1986). Each will be discussed in turn.

Microwave radiometry attempts to discern differences in tissue temperature through differences in microwave energy transmitted from the tissue to a receiver. As

such, it is a passive detection system similar to infrared detection. However, it differs in that microwaves (λ about 1-10 cm) are capable of penetrating several cm of tissue while infrared waves (λ about 0.01 cm) are essentially detectable only from the skin. Microwave imaging typically uses 2 or more frequencies in the range of 1-6 GHz. It is desirable to map the internal temperature of the body for several reasons. Inflammations and infections cause local heating, and there is no good way to detect the heat (besides invasive physical thermometers) if they are too far under the skin. Also, in hyperthermic treatments, a cancerous volume of tissue is heated via microwaves to a high temperature in an attempt to kill the cancerous tissue without damaging the healthy surrounding tissue. The cancerous portion must be held to within about 0.5 C of 43 C for about 45 minutes (Land, 1995). Presently invasive "meat" thermometer probes are used, and their numbers and depth of penetration are limited by trauma considerations, so that the thermometry of the area is usually inadequate. A noninvasive microwave method that could scan and heat the entire affected area is most desirable.

Microwave radiometry depends on complicated modeling and calibration (Mazushina et al. 1995) and as such is a difficult problem that still needs much development to become useful and reliable. Also, the amount of microwave energy emitted by the human body is extremely small, making it difficult to get good signal-to-noise ratios. As a passive method it is not of much interest to us, but is mentioned as a use of microwave radiation emitted from the body itself.

The study of chest motion, due to the action of the heart and lungs, has been examined using Doppler-shift radars where the frequency of the reflected signal shifts due to the relative motion of the reflecting chest-lung wall. The frequency change is

difficult to detect as the motions are not large (sub-cm for movements of the chest wall due to heart motion) and the velocities are very small, resulting in tiny Doppler frequency shifts. Any relative motion of the subject can completely submerge the signal. Chan and Lin (1967) used a 10.5 GHz X-band Doppler transceiver located at about 3 cm from the chest wall with some success, but the sensitivity both to relative motion of the subject and antenna and physical placement of the antenna precluded any practical use.

The imaging of tissue using microwaves has been referred to as "radar tomography". It is perhaps the most ambitious use of microwaves in the medical industry, and is promising and practicable if not yet practical. Both transmissive (or bistatic, where the subject is placed between transmitter and receiver) and reflective (monostatic, where the transmitter and receiver are located together) have been attempted. Bistatic (Lin 1986; Larsen and Jacobi, 1986; Meaney et al. 1996) applications suffer from multipath problems (the microwaves do not necessarily go straight through the tissue as X-rays will) and attenuation of the microwaves in the tissue. Monostatic (Murphy, 1994; Young and Peters, 1986) applications are also plagued with multipath problems and uncertainties due to scattering from multiple layers and the resultant interference. Both methods have problems with poor resolution at low (less than 10 GHz) frequencies, and poor penetration at high (greater than 10 GHz) frequencies. Yet the promise of non-ionizing, relatively inexpensive, real-time high resolution images from microwave interrogation spurs further research and one day microwave tomography may be as common as X-rays and MRIs.

However, for the moment, we are not concerned with complicated imaging problems. Our approach is simple – we just want to know how some tissues of the

trachea vibrate in response to vocal fold modulation of airflow. Our tool, the GEMS, is small, portable, safe, inexpensive, and ideally suited for small-amplitude motion detection. This is where the use of microwaves as a common biometric tool will begin.

2.2. The tissues of the vocal tract

It is important for the reader to understand the basic layout of the tissues of the vocal tract. In this chapter I will explain some simple anatomical terms and give an overview of the configuration of the vocal tract. As a picture is worth more than a thousand of my words, I have included a profile (Figures 2.15 and 2.16) of the tract from an excellent anatomical book, "Atlas of the Human Body", by Netter (1997). Additional information on any of the structures discussed may be found here as well as in Titze (1994).

Anatomical terms:

Anterior: toward the front

Posterior: toward the back

Superior: above

Inferior: below

It is also important to establish the position of the glottis, the air space between the folds. Many locations inside the tract are referenced from their position from the glottis:

Subglottal: below the glottis

Supraglottal: above the glottis

The vocal tract starts with the lungs, which act as a pressure or airflow source. They are usually modeled as a constant pressure source, easily accomplished with even flexure of the diaphragm. As phonation begins, lung pressure rapidly rises above atmospheric (but not far – even for loud voicing only about 1.016 atmospheres is reached), driving the air from the lungs through the trachea and into the midsection of the tract – the larynx.

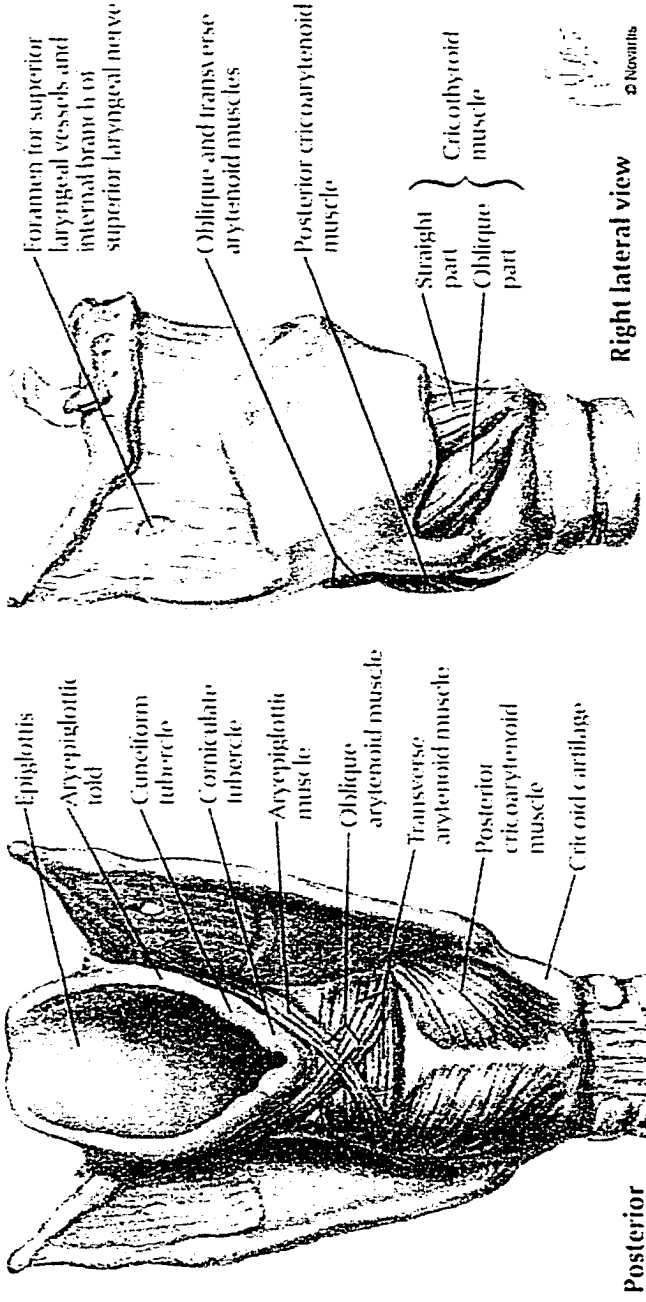


Figure 2.15. Posterior and right lateral view of laryngeal structures (Copyright 1997. Novartis. Reprinted with permission from the *Atlas of Human Anatomy*, illustrated by Frank H. Netter, M.D. All rights reserved)

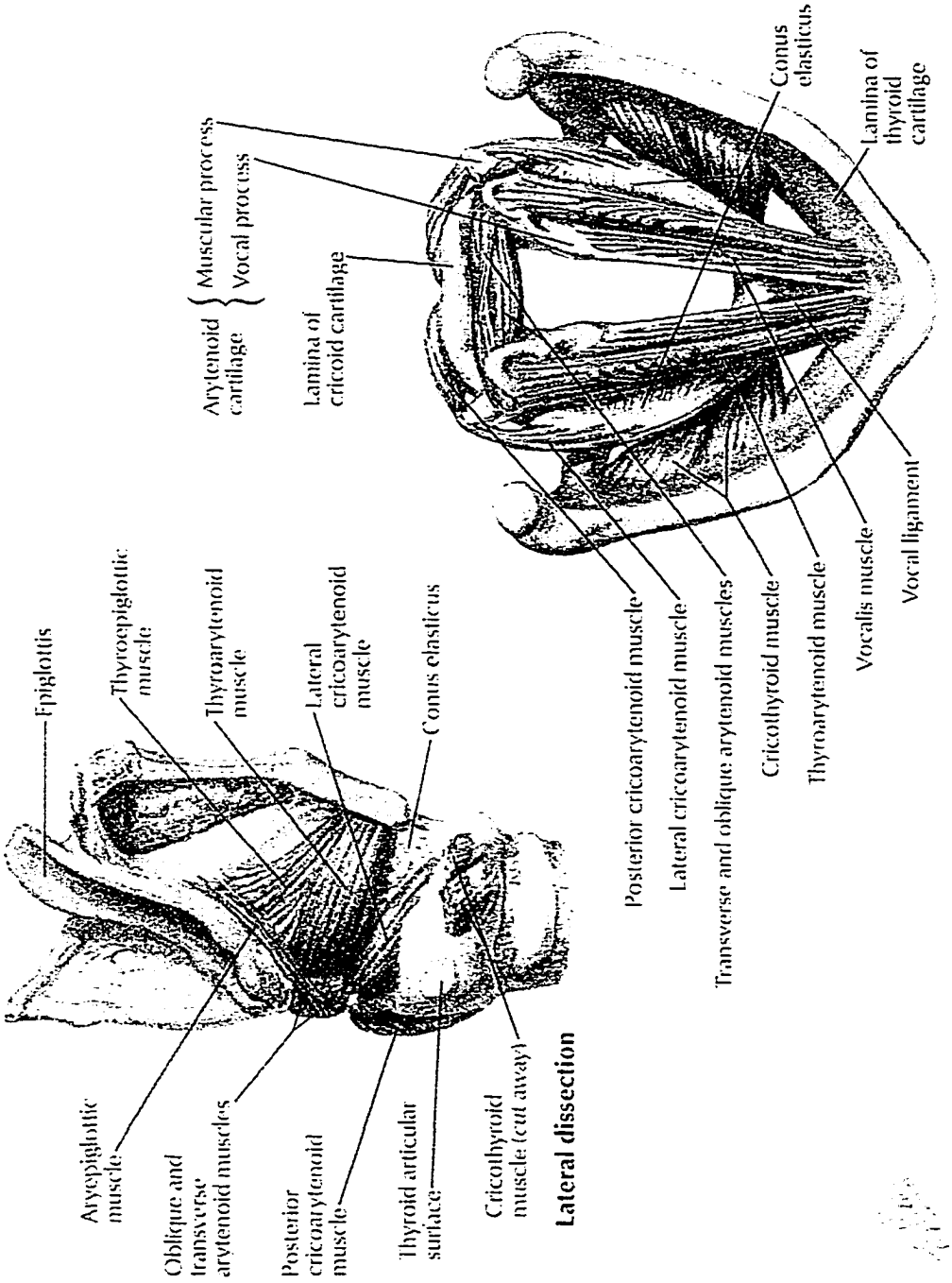


Figure 2.16. Superior and lateral dissection view of laryngeal structures (Copyright 1997. Novartis. Reprinted with permission from the *Atlas of Human Anatomy*, illustrated by Frank H. Netter, M.D. All rights reserved)

The larynx is considered to be the area from the superior tracheal rings to the epiglottis. It is a flexible structure that can move several centimeters up and down. It is composed mostly of cartilage and smooth tissue, including only a single bone (the hyoid) which itself floats with respect to the skeleton. Proceeding up from the tracheal rings we have the cricoid cartilage, the thyroid cartilage, the hyoid bone, and the epiglottis. The cricoid cartilage may be considered to be the most superior tracheal ring, but its construction varies significantly from the other tracheal rings. It is signet ring-shaped, being quite narrow in length anteriorly and quite wide posteriorly. Unlike all tracheal rings, it completely surrounds the trachea. The anterior portion (known as the "arch") can move up and down, and this is one method of increasing or decreasing tension in the folds. Attached to the superior posterior section of the cricoid cartilage are the arytenoid cartilages – small, tooth-shaped structures that are attached both to the vocal folds and a muscle that rotates the arytenoids, opening and closing the folds for breathing and phonating.

Next is the thyroid cartilage, the largest and most prominent of the vocal tract cartilage. The anterior-most portion of the thyroid cartilage is the laryngeal prominence, or "Adam's Apple". It is easy to determine the location of the prominence by feeling for the notch at the top and center of the thyroid cartilage. Directly behind this point is where the anterior portions of the folds connect to the cartilage. Thus the anterior portions of the folds are fixed, and the posterior portions are free to rotate on the arytenoids to allow for breathing, protection of the lungs from foreign substances, and phonating. Note also that the front and back of the folds are mounted on different

cartilage, allowing the folds to change tension by rotating the different cartilage with respect to one another. The folds themselves are discussed in Section 2.3.2.

Above the vocal folds and before the mouth and nasal cavities is the region known as the pharynx. Immediately following the vocal folds are the ventricular (or false) folds. These are a soft band of tissue that can, under some circumstances, press together and assume the role of the vocal folds. Usually this occurs when the vocal folds have been damaged through disease or overuse. The ventricular folds are a poor (but the only available) substitute for a speaker with damaged folds. The ventricular folds can also close over healthy folds, disrupting the flow of air through them and resulting in a raspy voice.

Next comes the epiglottis, a cartilage whose main function is to fold over the trachea and tightly seal it when food or water is being ingested. During breathing or phonation, it relaxes and forms a resonating chamber in the supraglottal region. As it is movable, it is one way that we are able to change the resonating qualities of our tract (by changing the cross-sectional area, see Section 2.3.1) in order to form different phonemes. The epiglottis is attached inferiorly to the thyroid cartilage and superiorly to the tongue.

Above the epiglottis are the openings to the oral and nasal tracts. The tongue is an important means of changing the cross-sectional area of the oral tract at different distances form the glottis. For instance, for the phoneme /i/ (see Appendix D for a list of American English phonemes and their symbols), the tongue creates a constriction at the front of the mouth, and for /u/ the constriction is near the back. The tongue is a marvelous instrument that allows many different phonemes to be formed as well as

producing clicks, pops, stop consonants (such as /d/ and /g/), and fricative consonants (/t/ and /θ/).

After the tongue are the teeth and lips. The teeth are an important part of forming speech, especially for fricative consonants, but they also contribute to the resonance qualities of the mouth. Without the teeth, air tends to bleed through the lips and give the speech a "sputtering" quality. The lips are another muscle group that aids in the shaping of the sound of the excitation function. They may be positioned in such a way as to lengthen (or shorten) the length of the vocal tract, lowering (or raising) the resonance frequencies of the tract. The jaw may be used in a similar manner. Lowering the jaw and widening the mouth causes the pressure in the mouth to drop closer to the glottis, shortening the effective length of the vocal tract tube and raising its resonance frequencies. This tactic is successfully employed during a scream to get the highest frequencies possible in the acoustic signal.

The nasal cavity is quite a bit more complex than the oral. We shall not go into much detail here, the interested reader may refer to Netter (1998). Guarding the entranceway to the nasal cavity is the soft palate or velum, at the end of which is the uvula, easily seen when peering into the oral tract. The velum is a tissue-covered cartilage, protects the nasal cavity from any food or water in the oral cavity, and in its spare time controls whether or not the nasal cavity is used as a resonator in the vocal tract. With the velum in the up position, the nasal cavity is not used. This is not, however, the case for most phonemes. When the nasal cavity is blocked or held shut the resulting audio is deprived of the resonances of the nasal tract and is said to sound "nasally", an ironic term since the nasal cavity is not used when this type of speech is

produced. Some phonemes (known as the nasals, /m/, /n/, /ng/) use only the nasal tract and pinch off the oral tract with either lips (/m/) or tongue (/n/, /ng/).

Above the velum is a series of interconnected nasal passages known as the sinuses. These sinuses eventually connect to the nose itself and thus the outside air. Each sinus cavity can act as a side branch of the vocal tract, generating anti-resonances or "zeros" in the spectrum of the acoustic output. These zeros absorb energy at and around a particular wavelength. They therefore remove energy from a spectrum, in contrast to resonances which reinforce the energy at certain wavelengths. This will be discussed at length in the next section. As far as adjustable mechanisms in the nasal cavity, there is very little a speaker can do consciously to change the qualities of the nasal tract. The walls of the tract can become coated with mucus due to inflammation from allergies or a cold, and this decreases the size of the cavities, changing the position of the zeros. Apart from this (and a little nasal flaring) the nasal cavity is relatively constant in size and acoustic response.

At this point is would be helpful to introduce two additional graphics from Netter (1997). In the first (Figure 2.17) a median section of the pharynx is presented. Notice that the cricoid cartilage is the only tracheal ring to completely encircle the trachea – all the others are U-shaped and do not make it all the way around. Also note that the posterior wall of the trachea is the anterior wall of the esophagus, and as such is quite flexible.

This tracheal feature is examined more closely in Figure 2.18. Here a cross-section of the folds is presented, along with a look at the longitudinal elastic fibers that make up the posterior wall of the trachea. The posterior surface, being flat with respect to the

trachea and composed of elastic fibers and muscle, would be an ideal surface from which to reflect electromagnetic waves. It is flat with respect to the anterior surface of the skin and it is composed of a high permittivity material (muscle and elastic fibers, which are also quite conductive). High values for the permittivity and the conductivity result in efficient scattering. It is also relatively thick (approximately 5-8 mm, about equal to the wavelength of the GEMS wave in the material), which would aid in reflecting the wave. So from an examination of the tissues, it appears that the subglottal posterior wall would be the most likely candidate for the origin of the signal.

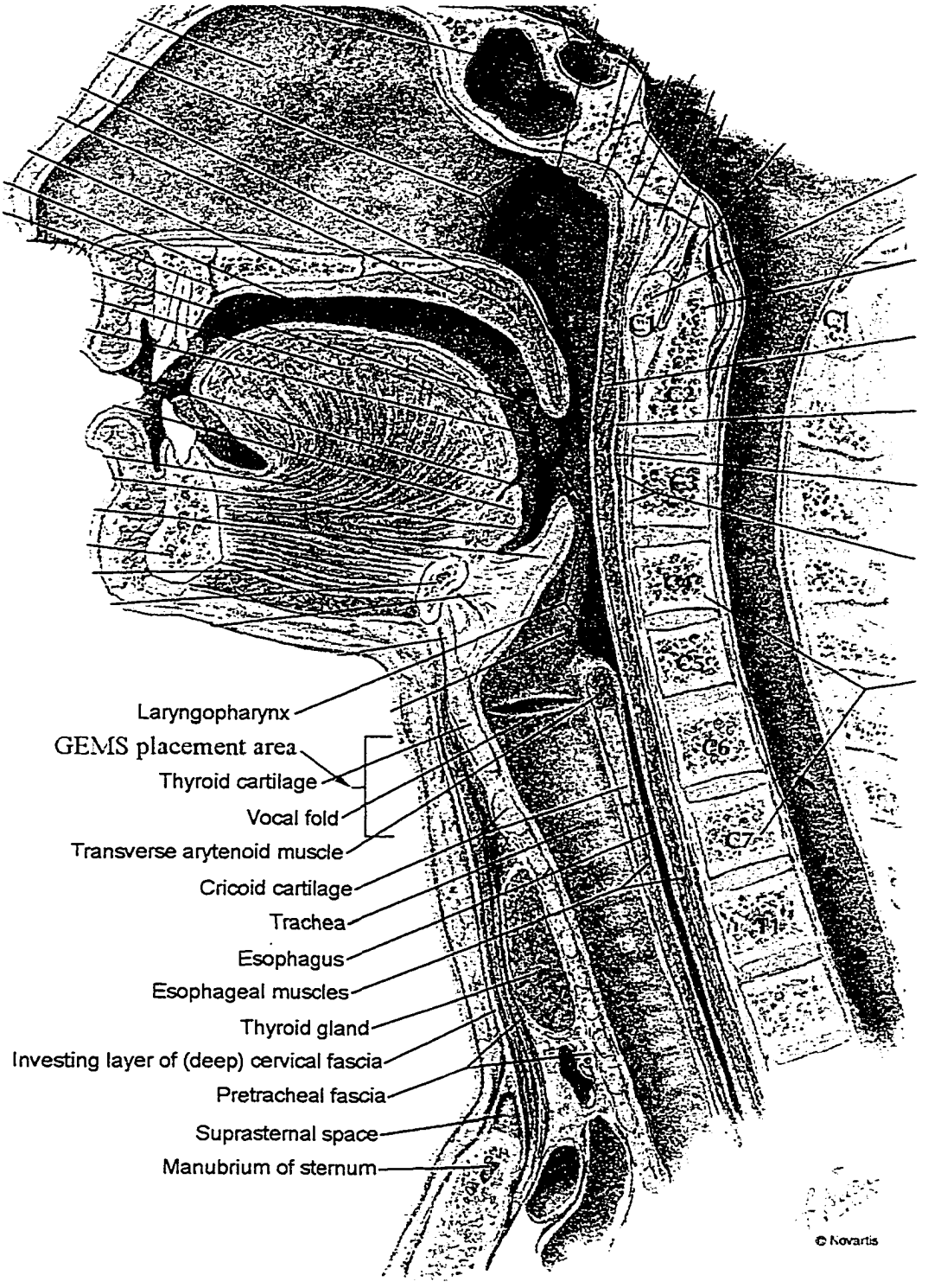


Figure 2.17. Median section of neck. (Copyright 1997. Novartis. Reprinted with permission from the *Atlas of Human Anatomy*, illustrated by Frank H. Netter, M.D. All rights reserved)

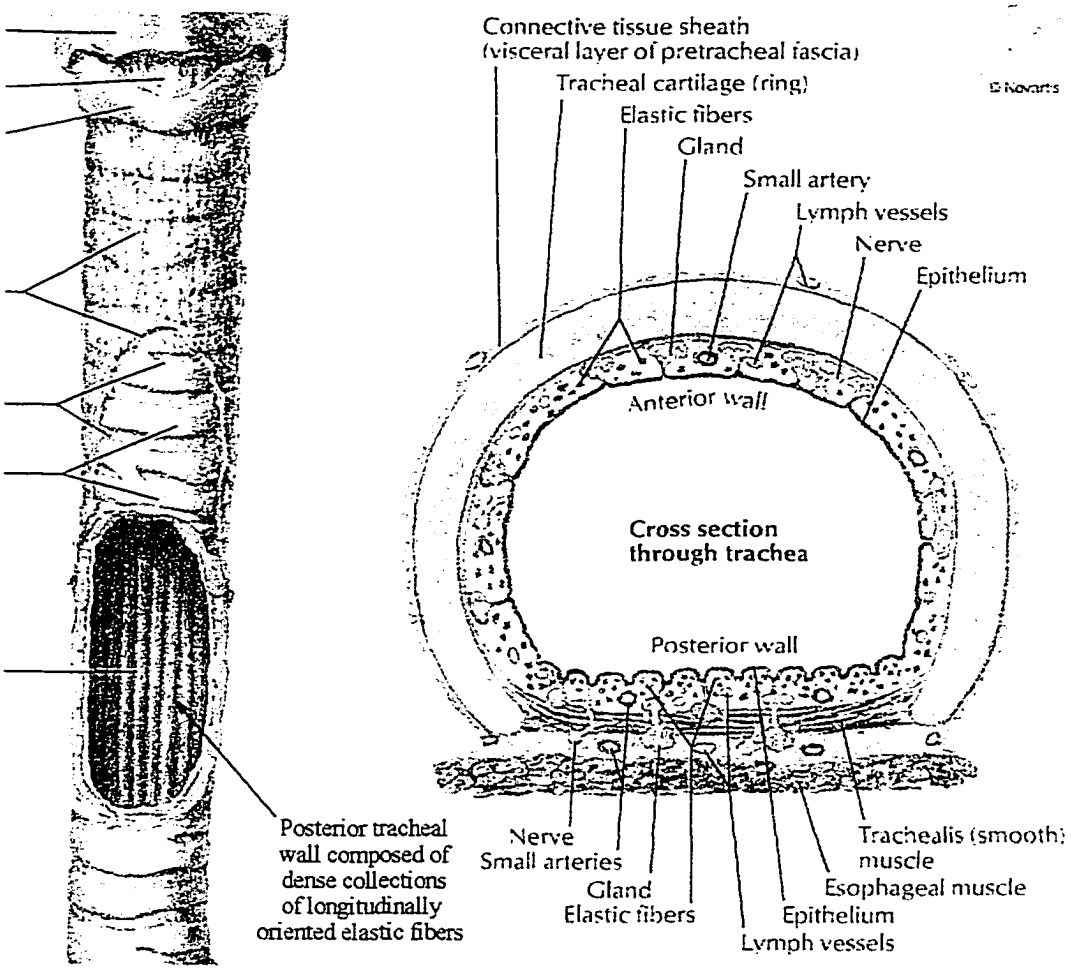


Figure 2.18. Cross-section of trachea. (Copyright 1997. Novartis. Reprinted with permission from the *Atlas of Human Anatomy*, illustrated by Frank H. Netter, M.D. All rights reserved)

Chapter 2.3. How we make and shape sound

The way humans form recognizable speech from a bit of air pressure and tissue motion is unique. We share many qualities with our animal brethren as far as producing sound is concerned, but none have the ability to phonate as long and steadily as we are capable of, and precise control of pitch is ours alone. In this section I will describe the basics of sound production in humans, along with models used to describe the vocal tract. I will also describe the methods used to calculate the transfer function of the vocal tract along with the assumptions involved. There is considerable detail about the mechanics of speech production that I have omitted in order to present an introduction – this is not a thesis about speech production *per se*. But it is important to have a grasp of the fundamentals, and understanding the resonance properties of tubes is vital in that understanding.

For those interested in more details, I recommend highly Dr. Titze's book "Principles of Voice Production". It is well written and Dr. Titze discusses many of the mechanisms of speech production.

Acoustic Impedance of Tubes

The transmission of sound requires a medium in which to propagate. Unlike transverse electromagnetic radiation, longitudinal sound radiation travels by disturbing the environment it is traveling through. The characteristics of that environment affect how the wave propagates: its speed, intensity, and frequency dispersion. Sound waves may travel through any media, but for our purposes we will restrict ourselves to the medium of air.

One of the characteristics of that medium is its impedance, its resistance to flow. Normally for unconfined air this is a constant value which depends on density and temperature:

$$Z_f = \rho\sqrt{\gamma RT} \quad \frac{kg}{sec \cdot m^2}$$

where Z_f is the free-space impedance, ρ is the density, γ is the adiabatic constant for air, R is the gas constant, and T is the temperature in Kelvin. When the airflow is confined to a tube, the impedance changes due to the effect of the tube. For example, the air particle velocity at the sides of the tube goes to zero due to frictional forces (Flanagan 1965), resulting in a stagnant boundary layer of air at the tube walls. Thus particle velocity is usually at a maximum at the center of the tube and it decreases as the wall is approached. Since the velocity of particles is not constant over the cross-section of the tube, the average particle velocity across a cross-sectional area perpendicular to the flow is used to define an impedance. The average velocity is multiplied by the cross-sectional area to get the average airflow in the tube. The average airflow (neglecting side branches and holes) is conserved in a tube if the cross-sectional area suddenly expands or contracts. For example, in a contraction of the tube area, the airflow remains the same, so the average velocity must increase. This is a simplistic (but effective) approach to understanding Bernoulli's law, based on the conservation of energy (which we have assumed). To define the impedance of a tube, then, we simply take the ratio of acoustic pressure p to airflow u (where u is the per-unit-volume mass flow):

$$Z_t = \frac{p}{u}$$

If there is travel in one direction only (1-D approximation, no reflections) this may be expressed as (Titze 1994 p. 137)

$$Z_t = \frac{\rho c}{A} \frac{kg}{sec \cdot m^4}$$

where c is the speed of sound and A is the cross-sectional area. In this simple form the impedance of the tube is the free-air impedance divided by the cross-sectional area. Thus any change in the cross-sectional area of a tube will lead to a change in impedance of the medium. This gives rise to reflections at the boundary between one impedance (cross-sectional area) and the next.

The reflection coefficient at the boundary is defined as the ratio of reflected pressure to incident pressure:

$$R = \frac{p_r}{p_i}$$

Pressure and flow across the interface have to be continuous, and as a result the reflection coefficients can be represented in terms of the impedances (Kinsler and Frey, 1962):

$$R = \frac{z_2 - z_1}{z_2 + z_1}$$

or, using the above,

$$R = \frac{\rho_2 c_2 / A_2 - \rho_1 c_1 / A_1}{\rho_2 c_2 / A_2 + \rho_1 c_1 / A_1}.$$

If, as is the case in most instances, the density and velocity of the air is the same on both sides of the change of area,

$$R = \frac{A_1 - A_2}{A_1 + A_2}.$$

This result assumes the wave is traveling from tube 1 to tube 2. A negative reflection coefficient indicates a change in phase by 180° in the reflected wave. This occurs when the second tube is larger than the first – as in the open end of an organ pipe.

Therefore the reflection coefficient can be described by the difference in cross-sectional areas of the two tubes. This is an important result, as any arbitrary tube can be approximated by a series of finite tubes with discrete changes in cross-sectional area. This can be used to model the vocal tract, as we shall see.

Resonance condition for tubes

Tubes have specific excitation frequencies at which they will resonate. When this occurs, a specific set of frequencies experience constructive interference and all others cancel out. Tubes are excited by inserting audio signals or just blowing across an open end. In the first instance there is some amount of control over the frequency content of the excitation, but in the second the excitation is chaotic and somewhat "white" in frequency content (many frequencies are represented more or less equally). Therefore only a single frequency may be excited in the first case (resulting in a purer tone) but in the second many frequencies (or overtones) will be excited, resulting in a "fuller" tone.

So what is the condition for resonance? Let's look at two cases, shown in Figure 2.19: First, the open-ended tube, then the closed-ended tube. Both are considered to be closed at the excitation end, although it is not necessary. A closed excitation end can be approximated by a small hole into which the signal is injected. An example is the small opening between the lips of a trumpet player.

Open-ended tube

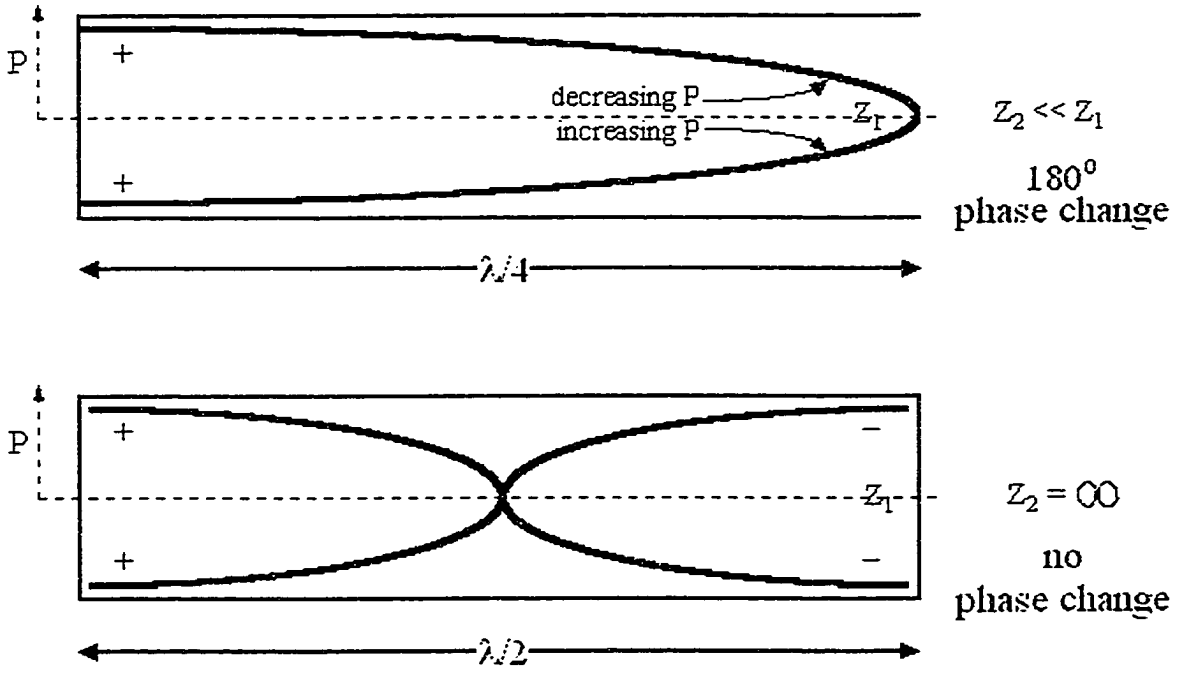


Figure 2.19. Standing pressure waves for the lowest resonances in an open-ended tube (top) and close-ended tube (bottom)

For the open-ended tube, the reflection coefficient R approaches -1, as region 2 is simply open air and $A_2 \rightarrow \infty$. Therefore the incident waves are almost totally reflected and out of phase with the incident waves by 180°. The frequencies that will experience constructive interference are those with a wavelength such that they will travel to the open end, reflect and shift by 180°, then travel back to the closed end. At this point constructive interference will occur if the phase of the reflected pressure wave matches the incoming one. Or, put another way, we want the phase of the wave after the second reflection to be the same as the incoming wave so that they will add constructively. The wavelengths that satisfy this condition for a tube of length *l* are those with a total travel length of $n\lambda = 2l + \lambda/2$, or $\lambda_n = \frac{2l}{n - \frac{1}{2}} = \frac{4l}{2n-1}$ (see Figure 2.19). In terms of frequency, $f_n = \frac{(2n-1)c}{4l}$. Therefore, the fundamental frequency would occur at n = 1,

with overtones present for higher n. At this point I will introduce the term *formant*, which the speech community uses in lieu of *resonance*. The first formant (for n = 1) is referred to as F_1, the second F_2, and so on. The open-ended tube is referred to as the quarter-wave tube, as its length is a quarter-wavelength of F_1.

Although the node at the open end of the tube in Figure 2.19 might look strange, it is correct for the following reason: The pressure at the closed end is not fixed. When a pressure wave propagates towards the closed end, it may compress the air there to higher densities than normal, as the air has nowhere to flow. As it reflects off the closed end and travels toward the open end, it will create a rarefaction. Thus the closed end contains an antinode (peak or trough) of pressure (it can vary all the way from high density to low). This might not be evident from the picture, but remember – the plot of pressure shown is only a snapshot of the pressure variation. Every antinode point will experience the full range of pressures, and every node will have its pressure value fixed. This is why the node is located at the open end – inside the tube the pressure may vary, but outside the tube the pressure is that of the surrounding atmosphere – which for all practical purposes may be considered fixed.

Closed end tube

At the end of a closed-end tube, $Z_2 \to \infty$, so $R \to 1$, and there is no phase change at the boundary. Thus the travel distance is now $n\lambda = 2l$, leading to formants at $f_n = \frac{nc}{2l}$. This is referred to as the half-wave tube. Here the node is located at the center of the tube for the lowest resonance. A tube with two ends that are open instead of closed has the same formant locations, but now the nodes are located at the open ends.

It is important to note that the closed end of the tube is only an approximation. If it were truly closed, no sound could escape from the tube! In practice there is an opening in the closed end, but it is quite small in comparison to the tube cross-sectional area. In the vocal tract, this is approximated by narrowing the lips. As a result of not being truly closed, the half-wave resonances shift downward toward the quarter-wave resonances. You can see the effect easily yourself. Begin by phonating a vowel (try "eh" or "oo"), and while keeping your pitch the same (not as easy as it sounds!) vary your lip opening from wide open to almost closed. You are in effect changing your vocal tract from an open to a closed ended tube and in the process raising the formant frequencies. This is one way we change the frequency content of the sound entering our vocal tract through the glottis.

Both of these types of tubes are used in a pipe organ in order to cover a wide variety of tones. Care must be made in the selection of length for frequency, however, as the actual location of the pressure node at an open end is not exactly at the end of the tube – it extends slightly into the surrounding air. For narrow cylindrical tubes, the length extension is about 0.6r, where r is the radius of the tube (Halliday *et al.* 1992). This leads to frequency errors on the order of a few percent, which are certainly noticeable and must be corrected for.

Frequency spectrum for tubes

The frequency spectrum of an object that simply modifies sound as opposed to generating it looks quite similar to the spectrum of a sound generator. It is important to keep in mind, however, that the tube (or filter) cannot add to the frequency content of the sound source; it can only modify some of the frequencies traveling through it. In a tube,

this is due to destructive interference of superimposed reflected pressure waves. You may inject white noise (equal amounts of every frequency) into a tube, but ideally the only frequencies that escape are at the resonances of the system. However, the resonances of the system often have some amount of width – they are not infinitely narrow. The width of the formants (spread in frequency from the ideal value) depends on the physical properties of the tube. Perfect formants are only possible if there are perfect reflections and no energy losses in the tube (which would require perfectly rigid walls). In the human vocal tract, naturally, sound is radiated from the mouth and nose and there is a good deal of flex in the tissues that compose the tract. Thus the interference (both destructive and constructive) is not as effective and formant widths are broadened. The width of the formants may be a good physiological marker for individuals (as they reflect the physiology of the tract), but not enough analysis has been accomplished at this time to determine if this is true. This will come to light in the speaker verification thesis to follow this one.

Multiple tube resonances

The vocal tract may be modeled as a series of tubes of differing cross-sections (Figure 2.20) where the relative cross-sectional areas of the segments determine the resonant frequencies of the tract. Briefly, each tube affects the formants of the other tubes in a predictable manner. For example, taking the simplest case of two tubes, the effects can be seen as follows. Let the two tubes be of equal cross-sectional area to begin. This combination has formant frequencies at values determined by the combined length and area of the tubes. As an example we'll say the first two formants are at 500 and 1500 Hz. If the area of tube 2 is now increased and that of tube 1 decreased, the

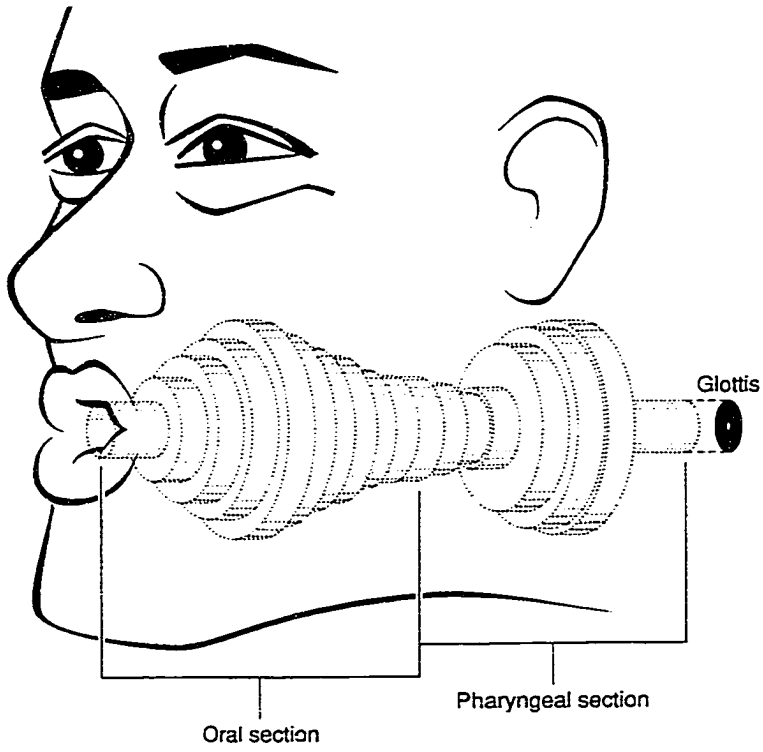


Figure 2.20. Cylindrical-tube approximation of the vocal tract for a simulated /u/ vowel (from Titze, *Principles of Voice Production*, 1994. All rights reserved. Reprinted by permission of Allyn & Bacon).

pressure drops in tube 2 and a "virtual node" appears in tube 2. This shortens the equivalent length of the tube and the first formant (F_1) is increased. At the same time, for the second formant (F_2) the node that would have been in tube one is "pulled" toward the virtual zero in tube 2. This increases the wavelength for F_2 and so it is decreased. So by simply tightening the "pharynx" (tube 1) and widening the "mouth" (tube 2), F_1 is increased from 500 to 700 Hz and F_2 drops to 1100 from 1500 Hz. This is precisely what happens when a speaker changes from the neutral phoneme /upside-down e/ (see Appendix D for a list of phonemes) to /a/.

The opposite effect occurs when the second tube is smaller then the first, as when an /i/ is phonated. The pressure increase in tube two causes F_1 to decrease as the mouth opening is quite small and the first formant slides back toward the closed mouth value.

F_2 in turn increases as the mouth tube appears to have two open ends, causing the nodes to migrate toward tube 2. Thus F_1 goes from 500 to 300 and F_2 increases from 1500 to 2300 Hz.

Similar effects involving lip rounding, three tube approximations (to account for rounding of the velum), and epiglottal motion lead to different locations of F_1 and F_2 for all the spoken vowels. There is some overlap (such as between /a/ and upside-down /c/) but most are distinct. In essence, we communicate by changing the shape of our vocal tract. As we change our pharynx diameter or tongue position we are modifying the frequency resonances of the vocal tract. That is why an /i/ sounds different than an /ɛ/. Different languages have different ways of modifying the tract and therefore have different phonemes.

The selection of the first two formants is not entirely just to make the example simpler. It has been shown (Peterson & Barney (1952), Bogert & Peterson (1957), and Kent (1978, 1979) that vowels are perceived and classified by the location of only the lowest two formants of the vocal tract (see Figure 2.21). The higher formants seem to carry other information about the characteristics of the individual's voice and may be where we get our ability to recognize speakers from even a short speech sample. So for simulation purposes, the two-tube model can be used to determine the gross vocal tract configuration required for phoneme recognition. If it were possible to reliably measure the locations of the first two formants in voiced speech, the phoneme being spoken could be identified with reasonable accuracy.

Hopefully by now the reader will understand the logic of trying to determine the cross-sectional area of the vocal tract. If it can be determined to sufficient accuracy, a

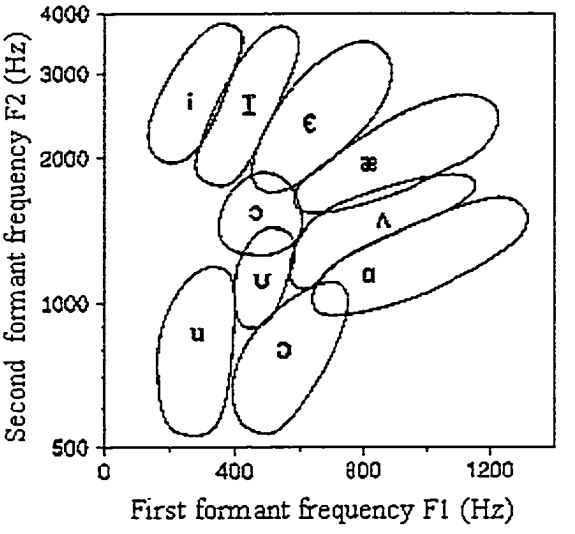


Figure 2.21. Vowel chart showing regions of F_1 and F_2 for 10 English vowels (from Titze, *Principles of Voice Production*, 1994. All rights reserved. Reprinted by permission of Allyn & Bacon).

model may be built to simulate the sound produced provided an excitation function can also be estimated. A variety of techniques have been used to measure the vocal tract area during phonation, including MRIs and x-rays, and these will be examined shortly. There are many difficulties inherent in these measurements. It is almost impossible to phonate for as long as it takes to get good scans, and the resolution of the systems used is still quite poor. Part of the advantage of the GEMS systems will be the sidestepping of this difficult procedure, in that the transfer function can be calculated using a GEMS-derived excitation function and the recorded audio. As a result we will have an excellent idea of the locations of the formants of the vocal tract, and in the future we may be able to infer the physiological processes associated with those formant locations. But for many applications this won't be necessary – once we have robust formant locations that is all we will need.

Measuring the physical dimensions of the tract

A variety of techniques has been used to measure or approximate the vocal tract area during phonation including electron beam computed tomography (EBCT, a form of x-rays, Story 1995), magnetic resonance imaging (MRI, Story (1995), Dang et al (1994), Sulter et al (1992), Moore et al (1992), Baer et al (1991)), x-rays (Fant (1960)), and cine-radiography (Moll (1965), a sort of x-ray movie, I hope the subjects were well compensated!). Molds of the vocal tract made from cadavers have also been examined, but the lack of phonetic tract configurations (currently only one) is limiting.

Of these, the only ones capable of volumetric imaging are MRI and EBCT. The EBCT (which has recently been shortened to just CT) scan is superior in resolution and it images teeth and bone, but as it utilizes ionizing radiation only a few scans per subject are considered safe. The MRI, on the other hand, is considered quite safe even for long exposures, but the acquisition time is at least an order of magnitude longer (~200 seconds per tract scan) and teeth and bone are poorly imaged due to their lack of water content. Over three minutes is a long time to phonate, so some inaccuracies due to subject movement are inevitable. X-rays are not used much anymore, especially in movie form due to the dangers they pose. Even where they are used (ECBT) they are renamed so as not to alarm the public.

There are many problems and difficulties inherent in measuring the vocal tract with the methods above. One is resolution – many use a scanning technique that produces "slices" of a particular resolution. On the present MRI systems, the slices are about 5 mm thick with each pixel in the slice representing about 1 cm^2. ECBT uses 3 mm slices and each pixel represents about 0.16 cm^2 (Story 1995). Even this leads to errors in

locating boundaries on the order of 2-3 mm, which can propagate and result in area calculation errors of 5-10% (Ibid.). The MRI area error rate is even higher.

The second is safety - the subject in Story's (Ibid.) experiments was exposed to more than two rads of radiation for only two volumetric EBCT scans. Two scans is not sufficient to quantify the vocal tract volume for normal speech, but the radiation dose is too high to do more. The dangers from repeated X-ray exposure are well known. The MRI has no known adverse effects, but is limited in resolution as stated above.

The third is lack of generality – these volume determinations can only be done in a closely controlled environment for a few individuals, and are quite time consuming. It is not something that is easy or inexpensive to do.

Thus, the present methods of calculating the cross-sectional area of the vocal tract are lacking in one or more respects. They are either not accurate enough or too dangerous. Therefore the use of the GEMS to calculate the resonances of the vocal tract (as discussed in Section 2.4), is advantageous as it is benign and negates the need for invasive or dangerous physical measurements.

Chapter 2.3.2 Sources of sound in the vocal tract

Now that we have explored cylindrical tube resonances and how they may be applied in an approximation to the vocal tract, we turn to the excitation function of the system. What are the excitations of the vocal tract?

Everyone is familiar with the vocal "cords". They produce the sounds that constitute our voiced speech. They are referred to in physiology as the vocal folds, as it is more descriptive of their appearance and function. The vocal fold's primary responsibility is to keep everything (except air) out of your trachea and lungs. They do this very well, as I found out during an experiment with a pressure catheter that was lowered between my folds into the subglottal area. The folds react quite strongly and involuntarily to contact with physical objects. The folds are essentially two floppy slabs of tissue composed of three main layers (see Figure 2.22) - the mucosa (composed of the epithelium and the superficial sublayers), ligament, and muscle. The muscle is the deepest layer and allows the tension of the folds to be changed so as to allow different frequencies of vibration. It is about 7 mm thick. The ligament is 1-2 mm thick and composed of elastin fibers (and some collagen) that allow for sustained phonation. The fibers are oriented longitudinally, like bundles of rubber bands. They give the "cords" their name, and only in humans are they a significant component of the folds. This is the layer that allows us to sing and speak continuously. The mucosa is about 0.5 mm thick and is also composed mainly of elastin fibers, but they are more loosely organized and flexible. This layer is more akin to a water-filled balloon than a rubber band.

The folds retract for breathing but can seal the trachea completely (and quite involuntarily) so that fluids cannot pass. When they are not busy saving us from

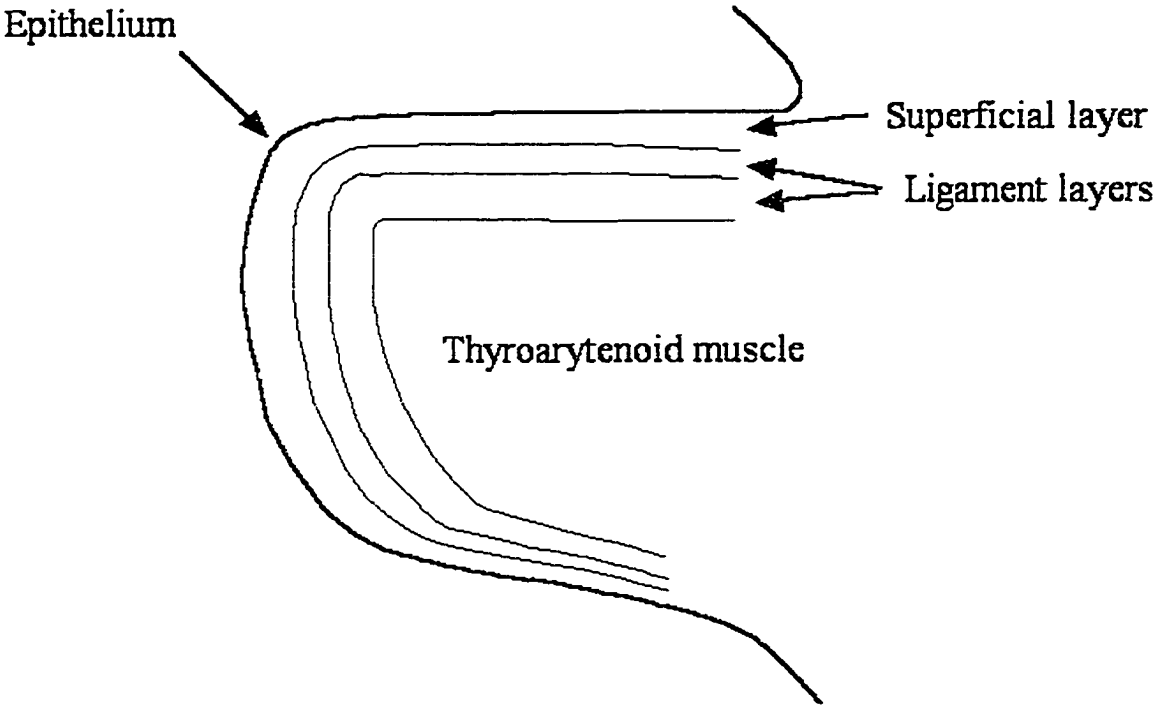


Figure 2.22. Sagittal (front to back) cross section of a vocal fold.

drowning, they are capable of relaxing slightly and allowing "bubbles" of air through. When the pressure is high enough and the folds under the correct tension, a self-sustaining oscillation can occur. This oscillation causes airflow and pressures pulses to propagate through the vocal tract and out of the mouth. Some will be reflected in the tract and reflect off the folds again, so the oscillation of the folds is somewhat dependent upon the reflected wave. However, the effect is small and (in the best physics tradition) shall be neglected in this introduction.

One-mass model of vibration

In order to be self-oscillating, the net energy loss per cycle has to approach zero. As there are many losses associated with friction, heat, and acoustics, there must be some manner in which energy is added to keep the cycle going. The two things that are required in order for this to occur are: a tube above the folds, and a pliable cover that can support wave motion.

To explain this, I will introduce the one-mass model of fold vibration (Titze 1994, see Figure 2.23). In this model the folds are a single rectangular block of tissue which oscillates perpendicular to the tract. It is considered to be a simple harmonic oscillator and has mass m, stiffness k (effective stiffness of the fold layers), and damping constant b (representing the viscosity of the tissue).

The intraglottal pressure P always acts perpendicular to the fold surface. If P varied depending on the direction of tissue motion (specifically if it were greater during opening than closing), it would be able to give energy to the folds and sustain their oscillation despite the damping. But how could this occur?

A derivation of P using Bernoulli's energy law can be found in the literature (Titze 1983, 1988). A simplified version, with no fold collision and ideal flow through the glottis, is

$$P = (1 - a_2/a_1)\cdot(P_s - P_i) + P_i$$

where a_1 and a_2 are the cross-sectional areas at the glottis entry and exit, respectively. P_s is the subglottal pressure, and P_i the input pressure to the vocal tract. Thus the intraglottal pressure P depends on both the transglottal $(P_s - P_i)$ pressure and the supraglottal pressure P_i. For the single-mass model, $a_1 = a_2$ and P is simply equal to P_i. Thus, something must happen above the glottis in order for P to vary with fold motion.

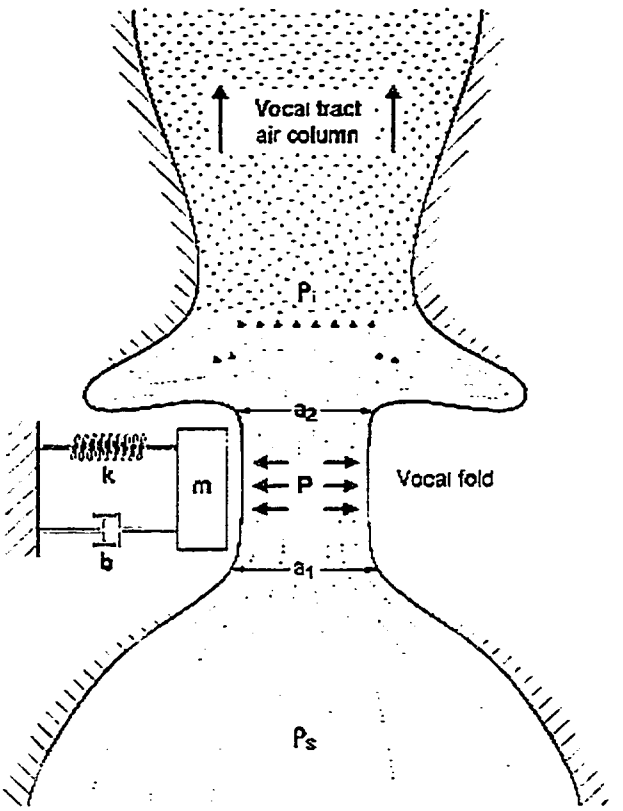


Figure 2.23. A one-mass model of the vocal folds, including airflow through the glottis. pressure against the tissue wall, and a supraglottal air column (from Titze , *Principles of Voice Production*, 1994. All rights reserved. Reprinted by permission of Allyn & Bacon).

What occurs is this (from Titze 1994, Flanagan and Landgraf (1968)): As the folds open, airflow is accelerated through the glottis and into the vocal tract. This influx of air causes the supraglottal pressure P_i to rise. In this approximation, then, the intraglottal pressure P also rises, forcing the folds apart. The increasing flow adds momentum to the air column above the glottis. As the folds reach the point of maximum flow (maximum glottal area) and rebound, flow begins to decrease. However, the column of air above the glottis continues to move forward. It has inertia, and the decreased flow through the glottis cannot keep up with the movement of air. Thus a negative pressure (suction) is produced at the superior surface of the folds, pulling them together. The key element is the inertia of the confined flow of air above the glottis. Without it there would be no

restoring force on the open folds and no sustained oscillation. With it, P_i helps to drive the folds in concert with their natural motion.

In more quantitative means, consider the analogy between the moving air column and a moving mass. Pressure is analogous to force and inertance corresponds to mass. Inertance is defined as

$$I = \frac{\rho L}{A}$$

where ρ is the density of the air, L is the length of the column of air, and A is its cross-sectional area. In accordance with Newton's second law, then,

$$P = I\frac{dU}{dt}$$

where U is the volume flow (in m^3/sec) of the air column. The volume acceleration of the air column is dU/dt.

In this case, then, the pressure is positive if the acceleration is positive, as it is at the beginning of flow through the glottis. The pressure is negative when the column begins to decelerate during the closing phase of the folds. Thus, energy can be transferred to the folds by the supraglottal pressure, which varies in rhythm with fold motion.

It is important to note that the tracheal tissue surrounding the enclosed P_s is also flexible and will be affected by changes in P_s. As we saw in Section 2.2, the rear tracheal wall has no cartilage support and is composed of muscle and elastin tissues, allowing it to move with the changes in pressure.

This explanation may sound like just the Bernoulli effect in action, but it really depends on the column of air in the vocal tract. The Bernoulli effect on the fluid as it passes through the glottis is always to decrease the pressure due to the increase in

velocity as it passes through the constricting glottis – it is insensitive to the direction of motion of the folds. Another force is needed to counteract this effect in order that the folds may open. The column of air supplies the positive pressure needed to help open the folds as well as the negative pressure to close them.

There have been a number of more complicated multimass models postulated (see Gauffin et al. 1983, Scherer and Titze, 1983, Ishazaka and Matsudaira (1972), Titze (1988), McGowan (1991), and Titze (1994)), but these will not be discussed here. The basic mechanics remain the same, except multiple modes of vibration are possible and a modeled mucosal cover adds to the pressure effect of the superior column of air.

Registers of voiced speech

During spoken speech, the folds may oscillate in one of three main modes or registers. These are referred to as modal (or chest), falsetto, and fry (or pulse) (Titze 1994). Slightly different registers are used for singing, but we will not concern ourselves with them. The registers are classified on a physiological basis, but as it happens they differ in frequency, so we will examine them in that manner from lowest to highest.

Vocal fry is a phenomenon that is quite common and sounds like a set of pulses rather than a single tone. It occurs when there is one of two things present: not enough lung pressure to set up consistent oscillations, and/or a physical asymmetry in the folds where one is larger than another. As a result of the first condition, instead of harmonic oscillations of the folds, inconsistent "burbling" of air through the folds occurs, much like the way steam burbles through mud in the fumaroles at Yellowstone Park. The excitation from one bubble of air dies out before the next can begin and the listener perceives the sound not as a continuous tone, but as a pulsed one.

The second condition of asymmetric folds is different in that the oscillation is consistent, but the pulses caused by closure are not consistent in amplitude. What normally happens is that the pulses alternate in amplitude, leading to a subharmonic at one-half the oscillation frequency (Ibid. p. 259) due to amplitude modulation. A voice with this quality sounds more "raspy" than the one with insufficient lung pressure and can be the sign of several disorders. A subject with a disorder of this type was studied in Section 3.3

Modal or chest register is the normal register for most people. It is referred to as chest because the chest can be felt vibrating during speech.

Falsetto is familiar to those who enjoy Hawaiian and Western music, in which it plays a large role. In this mode, the thyroarytenoid muscle contracts, bulging out the mucosa. The mucosa then acts as the primary vibrator as opposed to the ligament during modal speech. On occasion the folds may pinch together, reducing the length of the vibrating mucosa and thereby raising the pitch of the sound. The folds may even part enough so that they are not touching, but are still modulating the airflow through the glottis.

Unvoiced excitations

Vibrating folds are not the sole way we produce recognizable sound. Whenever we whisper, we are creating turbulent air and using the noise created as the source in our vocal tract system. The sounds may be produced anywhere along the tract: from the glottis (whispering), the back of the tongue (/k/), the tip of the tongue (/s/), or the lips and teeth (/f/). If from the glottis, it is termed *whispering*, otherwise it is known as *aspiration*

and can occur along with voiced speech. Even at the glottis, poor closure (leading to a good amount of "blow-by") can lead to a "breathy" voice.

In addition to sounds produced by turbulence, there are transitory sounds that are the result of the sudden onset or offset of phonation. These are referred to as *glottal clicks* and *glottal stops*. These include /b/, /p/, /t/, and /g/.

Lastly, sounds can be produced by the movement of liquids on or near the vocal folds. This lends a "gurgling" quality to the sound and can be the sign of a disorder if excessive.

2.4 Propagation of sound through the vocal tract and skin

We have explored the way sound is produced in the tract as well as the effect a tube has on the frequency content of an excitation. Now we attempt to bring both together and examine the way the sound propagates through the vocal tract and the surrounding tissues. For the moment we restrict ourselves to those sounds produced by the vocal folds in motion – voiced speech. The propagation for the other types of sounds discussed above is similar, only at times the entire length of the tract is not used.

Overview

As the lungs pressurize and force air through the glottis, sound (in the form of pressure changes or equivalently, airflow pulses) is produced. Due to the oscillation of the folds, the constant pressure in the lungs is modified into pressure pulses at the glottis, which propagate both in the subglottal and supraglottal directions. When the folds are closed, the subglottal pressure builds and the surrounding subglottal tissues swell in response. These tissues store energy in the form of elastic tissue deformation. When the folds begin to open, the stored energy in the subglottal region is released, albeit not all at once as the resistance of the folds is still quite high when the glottis is small (see Figure 2.24, derived from values given in Flanagan (1965), p.40). As opening of the folds progresses, the excess pressure in the subglottal region is relieved and the subglottal tissues rebound and even get pulled in slightly by the Bernoulli force of the nearby flowing air. As the glottis reaches its maximum size and begins to close, the impedance of the glottis increases, but only slowly. At this point the subglottal tissues are quite relaxed and flexible. As the glottis closes, resistance increases dramatically and the subglottal tissue expansion is the only outlet for the now-capped airflow. The tissues fill

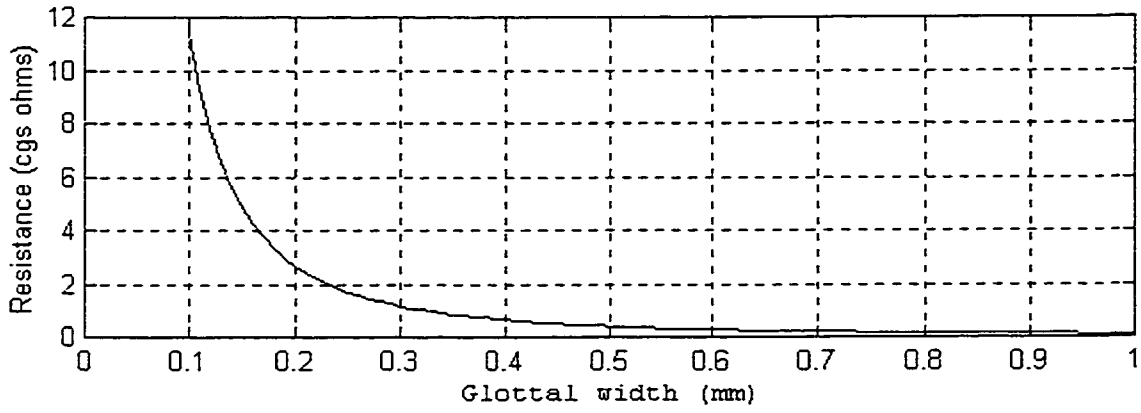


Figure 2.24. Glottal resistance for moist, warm, viscous air vs. glottal width (assuming glottis is rectangular)

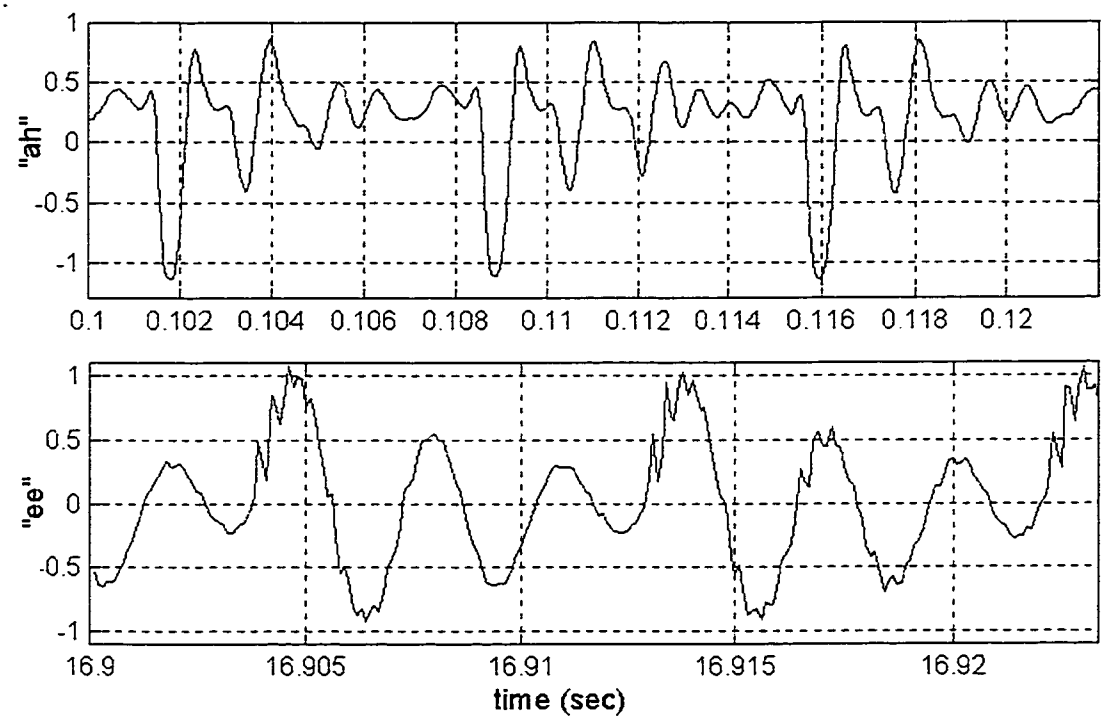


Figure 2.25. Normalized audio traces for /a/ "ah" (top) and /i/ "ee" (bottom). The duration of both are 22 msec. Note how the "ee", although longer in period, has more amplitude than the "ah", which loses energy more quickly.

with air suddenly, like wind snapping a limp sail into its full breadth. Supraglottaly, the closure of the folds results in a pulse of airflow into the supraglottal vocal tract. A similar (but opposite in sign) pulse is generated in the subglottal region, but the lungs are

quite lossy so little reflection occurs. In both the sub- and supraglottal tract, the airflow reflects from the many different cross-sectional area changes in the vocal tract (the largest normally being the mouth and nose openings, the subglottal tract is relatively uniform until the lung is reached). At each interface, the reflection is not perfect and thus some energy is transmitted past the interface and into the subsequent section. Some energy is also lost into the surrounding tissue in the form of heat. Therefore, after glottal closure there is a "ringing" of pressure waves in the vocal tract which eventually dies out. How much dies out before the next cycle begins is a function of the phoneme being spoken and the pitch of the vocalization. For example (see the plot of normalized audio in Figure 2.25), with /a/ the loss in energy (amplitude) is quite rapid, but with /i/ it is much slower. The difference is even larger when the relative periods are taken into account. The period of this /i/ is 9.1 msec (pitch of 110 Hz) while the /a/ is only 7.2 msec (140 Hz). Thus even with an extra 2 msec between impulses the /i/ retains more of its energy. This is due to the constriction caused by the positioning of the tongue near the roof of the mouth during the voicing of an /i/, leading to a higher reflection coefficient at that point. As a result less energy travels to the mouth and is radiated. In contrast, for an /a/ the mouth is lowered significantly (as doctors have noticed) and the impedance at the tongue point is quite low, so most of the energy escapes and is detected by the microphone in a few milliseconds.

After closure and excitation of the tract, the process repeats – the folds oscillating at their own frequency, the sub- and supraglottal tissues of the trachea following in concert. This vibration may be felt outside the skin just under the laryngeal prominence (the so-called "Adam's Apple"). Since the subglottal tissues are closely correlated with the

vibration of the folds, they are capable of yielding information concerning the pressure variations caused by the fold action. They are not perfect, however – as a driven system there is a finite amount of time required to change frequencies of vibration. These are noticeable when the mode of vibration of the folds is quickly changed – such as the change from modal to falsetto and back.

As a result of the proximity of the vibrating subglottal tissue to the surface of the skin, the skin is moved and sound is radiated from the trachea. It is easy to reproduce – simply close your mouth and nose and try to speak. Sound cannot escape through your nose and mouth, but it is radiated from your trachea until the vocal tract pressurizes to the point air can no longer flow through the glottis.

Much research has been done in the area of subglottal pressure measurements and approximations, but little has been done on the effects of the pressure on the surrounding tissue. One noteworthy exception is Ishizaka, K., French, J.C., and Flanagan, J.L. (1975). In this paper, the authors directly determine the impedance of the neck using a mechanical shaker and accelerometer. The values obtained from this work will be used in Section 4.1, where we will describe a process that removes the effects of the vibrating tracheal wall from the GEMS return, leaving us with the original driving subglottal pressure.

2.4.1 Lumped-element circuit models

The vocal tract may be approximated by a series of concentric right circular cylinders with varying cross-sectional areas, as discussed in Section 2.3.1. These cylinders may in turn be modeled by electrical circuits, in which pressure takes on the

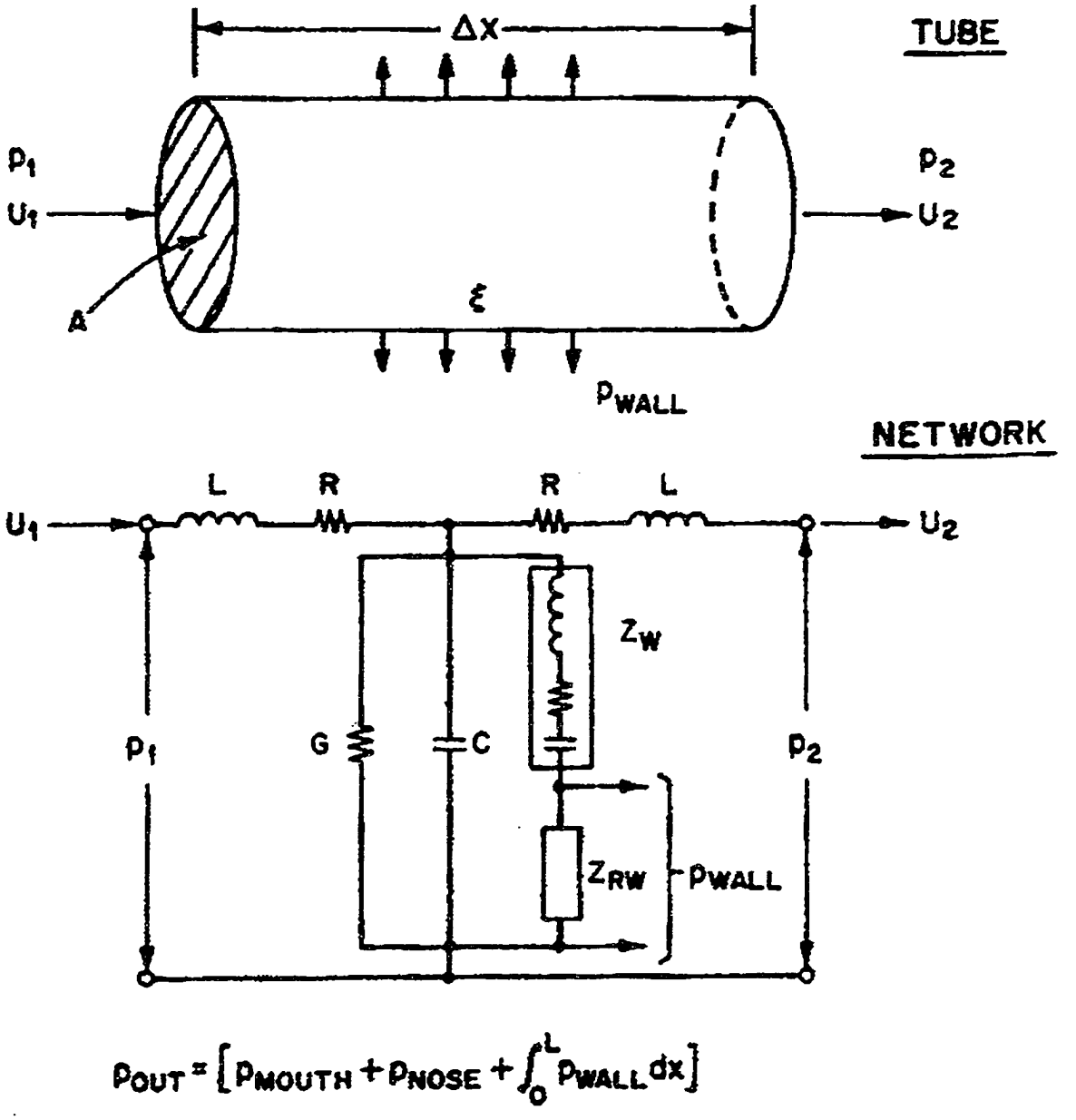


Figure 2.26. Equivalent circuit for plane acoustic wave propagation in an incremental yielding tube (from Ishizaka, French, and Flanagan (1975)).

The relation between physical parameters and circuit values are given in the lower part of Figure 2.26. L is the per-unit-length inertance of the air mass of the cylinder element, R is the viscous loss at the sidewall, G represents the conductance of heat into the sidewall, C is the acoustic compliance of the contained air volume, Z_w the equivalent mechanical impedance of the yielding sidewall, and Z_{rw} the radiation impedance of the wall, assumed to be that for a pulsating right circular cylinder. Once suitable values for the above can be deduced or measured, models of good accuracy can be built (either actual circuits or simulated ones). When combined with an excitation function, synthesis of speech can be performed, although for physical models only a single phoneme at a time can be simulated. The difficulties in using this model lie in getting good estimates for the circuit components, deciding on how many cylinders to use and their respective cross-sectional areas, and the synthesis of a realistic excitation function. Nevertheless, many authors (Story and Titze (1996), Story (1995), Narayanan *et al.* (1995), Sulter *et al.* (1992), Moore (1992), Baer *et al.* (1991), Flanagan *et al.* (1975)), have had considerable success synthesizing single phonemes, especially in a singing voice (simpler to synthesize as it is relatively constant in pitch and intensity). This success has not been easily obtained and has involved complicated time-consuming measurements of vocal tract area functions using x-rays and MRI images on live volunteers, and plaster castings and physical measurements on cadavers. The resulting calculations are cumbersome and lengthy, and oftentimes the results are only for a few phonemes, with transitions between them neglected. A more convenient method is needed in order to fully parameterize a talking human.

2.4.2 Signal Processing Methods

Another new and different method, made possible by the GEMS, involves the use of signal processing to determine the characteristics of the vocal tract. This bypasses the difficult direct measurements of the vocal tract, but can yield "nonphysical" results as it uses the input and output of the entire vocal tract, not just the larynx, pharynx, and oral tract. It is possible to account for all sources of sound (including the nasal cavities and skin radiation) using signal processing, and the resulting system does not necessarily have to be related to a physical tube if it accurately describes the characteristics of the entire vocal tract.

We shall concern ourselves with linear time-invariant (LTI) systems. These are systems that have a linear response to an input and do not change when the input is shifted in time. These systems are well understood and form the basis for electronic theory. Naturally LTI systems are only an approximation to the complex and nonlinear vocal tract, but the nonlinearities are assumed small and neglected for the present discussion.

LTI systems can be modeled as shown in Figure 2.27. We may operate in either discrete time (top of arrows) or discrete frequency (by a z-transform of the time responses, in arrows). On the left we have the input to the system x[n]. It is fed into the plant, which can be a simple filter or a complex group of control systems. The plant modifies the signal in a linear, time-invariant manner and yields the output y[n]. The plant may be described either by its impulse response h[n], or by its transfer function, H(z). The transfer function is complex, with its magnitude describing the magnitude response of the plant and its angle the phase response: If

$$H(z) = x + iy$$

where x and y are functions of z, then the magnitude response is

$$|H(z)| = \sqrt{x^2 + y^2}$$

and the phase response

$$\angle H(z) = \text{atan}(y/x).$$

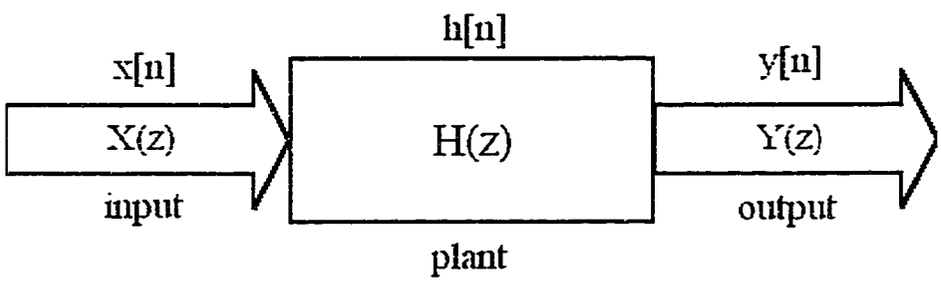


Figure 2.27. Schematic representation of an LTI system.

The transfer function describes the physical properties of the system. If you know the physical properties, you may deduce the transfer function. That is what has been attempted by many speech research teams, as described above. The other method (used in this paper) takes a different approach – the transfer function is deduced from the input and output of the system. That is, we have a "black box", a plant that we do not know the properties of. However, we measure both the input and output of the plant and from that we can calculate the transfer function. The accuracy of the transfer function calculation depends on the accuracy of the input and output measurements as well as the order of the modeling, as I will explain shortly. Thus it is imperative to get an accurate measurement of the input and output. The output of the vocal tract is naturally the spoken speech. The input, though, is more difficult to determine, and thus has never been used effectively in modeling the tract. The input to the system is either the acoustic glottal airflow or

equivalently the acoustic pressure injected into the vocal tract from the subglottal region. Before the GEMS, if an excitation has been used in a model of the tract it is strictly an estimation of the actual excitation. For the first time, the GEMS gives us the ability to measure the subglottal pressure which can be used as an excitation function. We finally will have the input to the system!

However, calculating the transfer function in this way does have a drawback: we may not be able to describe the physical system, as many different physical systems may yield identical transfer functions (known as multiplicity). But by utilizing physical constraints of the vocal tract system (the mouth can normally open just so wide, for example) it may be possible to eliminate most of the systems that yield the same transfer function and determine the real one. However, for the present work, knowledge of the actual physical basis of the transfer function is not necessary, as the transfer function itself supplies all the information needed for most speech processing.

An issue that may arise is that of time-invariance. How can we say the vocal tract is time-invariant when we are continuously changing its shape in order to speak the next phoneme? Also, the glottis is sometimes open and sometimes closed – how does that affect the model? Well, strictly speaking, the vocal tract is not time-invariant; it is a dynamic system. However, if the rate of change is slow enough, we may approximate it as a static system. That is one reason our algorithm normally processes two glottal cycles at a time. A variety of window lengths (from 1-10 glottal cycles) have been evaluated, and a length of 2 cycles has been found to give the best results. This is often enough so that the tract (and the pitch) changes little over this time period (10-20 msec), but not so

often that the sampled window has too few sample points. Too few points causes the Fourier transforms (and the resulting models) to be too crude.

With respect to the ever-changing glottis, in this paper the subglottal pressure is taken to be a good (if opposite polarity) approximation to the supraglottal pressure and therefore the input to the vocal tract. This is not strictly true after the glottis closes, as there is no excitation of the tract at that point and so the excitation should drop to zero. As we will see in Chapter 4, however, the calculated subglottal pressure falls rapidly to zero soon after closure and can therefore be used as an approximation to the pressure wave injected through the glottis into the vocal tract. Further refinements of this model are possible but will not be considered here.

Parametric models of the vocal tract

We now examine parametric models of the vocal tract. They are known as parametric since the transfer function is defined by a finite number of parameters (usually zeros and poles or their equivalent) rather than a continuous magnitude and phase response.

Why do we need a parametric model? Because of the tasks we wish to accomplish. We want to be able to enhance speech recognition and synthesis. For recognition, it is useful to know what phonetic units the speaker is forming over time. The best way to do that (and keep track of past phonetic output) is to use a compact model with a limited number of coefficients. That way the memory requirements don't grow too large for efficient operation. For synthesis, we want to be able to reproduce the speech from a specific human. The only way a computer can do that is with an excitation function and a

series of digital filters representing the change in phonetic output. Digital filters are best represented with a parametric model.

For any digital LTI system, the output y[n] can be represented as a linear combination of past inputs x[n-k] and past outputs y[n-k]:

$$A \cdot y[n] + B \cdot y[n-1] + C \cdot y[n-2] + \ldots = D \cdot x[n] + E \cdot x[n-1] + \ldots$$

where the y[n-k] are the k^{th} past output and the x[n-k] are the k^{th} past input. Transform now to frequency (z) space:

$$Y(z) \cdot \left(A + Bz^{-1} + Cz^{-1} + \ldots\right) = X(z) \cdot \left(D + Ez^{-1} + \ldots\right)$$

and since the transfer function H(z) = Y(z) / X(z)

$$H(z) = \frac{D + Ez^{-1} + \ldots}{A + Bz^{-1} + Cz^{-2} + \ldots}.$$

The roots of the numerator are the zeros of the transfer function (or plant) H(z) and the roots of the denominator its poles. For a further review of poles and zeros in the z-plane, please see Appendix A. Parametric models are more compact than nonparametric, but sacrifice a certain amount of flexibility. Ideally, any physical system could be realized with a finite number of poles and zeros, but in reality that is rarely the case as nonlinearities make some physical systems difficult to model. All the following models were implemented in Mathworks' Matlab 5.1.

ARMA

AutoRegressive Moving-Average (ARMA) models are those that incorporate both poles and zeros into the modeling process. This allows the modeling of both resonances and nulls of a system, but both the input and output of the system must be known. This is a superior model of the vocal tract because the vocal tract is not a simple tube, but has side branches that cause zeros in the transfer function of the tract. Remember that poles

are caused by the resonance frequencies of the various tubes that make up the tract, and zeros by "dead-end" tubes – those that are a branch of the vocal tract, but are closed on the other end. These "dead-end" tubes also resonate at certain frequencies, but as they are closed on the far end, the sound has nowhere to go and is eventually absorbed by the tissue. This has the effect of selectively removing certain frequencies from the vocal tract, creating "zeros" in the output spectrum. Most of the "zero tubes" are sinuses, both in the pharynx (the piriform sinuses) and in the nasal cavities.

An ARMA model is defined by the number of poles and zeros in the system along with any delays from the input to the output (in our case caused by the travel time of the audio wave from the glottis to the microphone). The coefficients and locations of each pole and zero are then calculated so the input will match the output when operated on by the plant (to within a set tolerance).

The principal difficulty inherent in parametric modeling is this: the numbers of poles, zeros, and delays of the model must be specified before processing takes place, and the number is not easily changed. As the vocal tract configuration changes for each new phoneme, the number of poles and zeros (while a good fit for some phonemes) may become inappropriate (the number of delays should not change appreciably as long as the mouth-microphone distance is relatively constant). Therefore it is good practice to use a few more poles and zeros than the physical situation might warrant to be sure to cover as many different vocal tract configurations as possible. A few extra poles and zeros will tend to cancel each other out (if there is enough of each). Too many poles and zeros (or some unmatched poles or zeros), on the other hand, can over-specify the system and begin to distort the transfer function. The number of poles and zeros we use to model the

vocal tract is continuously being refined, but at present is 14 poles and 14 zeros. That gives an adequate representation of most phonemes and is relatively stable over a wide variety of utterances and speakers.

LPC

Linear Predictive Coding is an all-pole model. It uses past values of the output to try to predict current values. It is useful where little or nothing is known about the input to the plant. In the equation above, the numerator would consist of a single coefficient (the gain). This is because there are no known past inputs, so all the coefficients on the right side (except for D, the DC gain) would be zero. LPC is specified by the order (number of poles) and the prefilter (the speech output is usually differentiated to emphasize the higher frequencies). LPC has been used for many years in the speech industry due to the lack of an input (i.e. an excitation function) to the vocal tract. It depends completely upon the audio (the output) to build an estimate of the transfer function of the system. As such, it is really an estimate of the spectral qualities of the output. It models the locations of the poles relatively well, but does not give any information on zeros. It is one reason present speech synthesis sounds like a person with a head cold – the zeros of the sinuses are not represented.

Cepstral

The cepstral is the odd child of the transfer function family. The cepstral method uses the Fourier transform (FT), but it is not a purely frequency based algorithm as it uses the FT twice to get back into a type of time space. The real (as opposed to complex) cepstral proceeds as follows:

$$X = \Im(x)$$

$$X_2 = \log_{10}(X \cdot X^*)$$

$$\text{Cepstrum} = \Im(X_2) \cdot \Im(X_2)^*$$

In essence, the real cepstrum is the power spectrum of the log of the power spectrum of the signal. The cepstral independent variable is known as the quefrency. The magnitude of the second power spectrum will have peaks at a quefrency time that correspond to repeated peaks in the first power spectrum. The first peak location is defined to be the pitch (Noll, 1966), and the information associated with the spectrum of the signal will be near the "DC" of the quefrency. To estimate the transfer function, only the first "few" coefficients are kept and the remainder zeroed out. "Few" is ill-defined and its selection is more art and experience than science, but perhaps the first 10% are sufficient. The truncated cepstrum is then inverse Fourier transformed to regain a power spectrum that has all the high frequency oscillations associated with the excitation function removed. This power spectrum is then used to approximate the transfer function. A property of the cepstrum is that the resonance peaks are quite broad, leading to decreased sensitivity to different speakers' idiosyncrasies. This is useful in speaker-independent recognition, but not in something like speaker validation. Also, the width can be broad enough so that significant overlap between formants can occur, leading to a decrease in accuracy.

Figure 2.28, constructed by Todd Gable, illustrates the differences in the three models. An artificial transfer function was constructed with two resonances and a single pole. It is displayed in the top trace, and the models built with ARMA, LPC, and cepstral algorithms are shown respectively below the actual transfer function. It is evident that for this system the ARMA model is superior, locating the resonances and zero almost

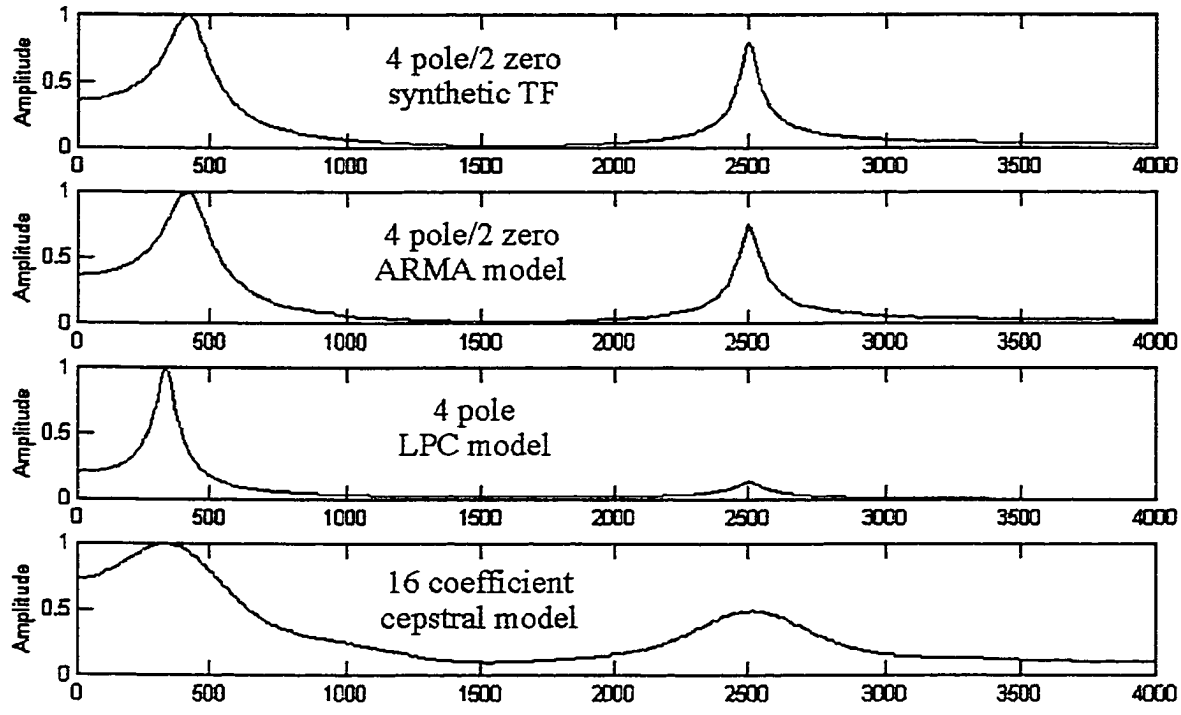


Figure 2.28. Comparison of a synthetic transfer function with 4 poles and two zeros (top) to three models: 4 pole/2 zero ARMA, 4 pole LPC, and 16 coefficient cepstral.

perfectly. The LPC model does not capture the zero, and the location of the first resonance is too low in frequency. The cepstral does a good job of capturing the features of the actual transfer function, but the excessive broadness of the peaks is quite evident. This demonstrates that if the system to be modeled has one or more zeros (as the vocal tract does), the ARMA model will do a better job of modeling it.

Thus the ability to use the ARMA model due to the information about the input gathered by the GEMS sensor should lead to more accurate and stable transfer functions that correspond reasonably well with the physical system they are representing.

Chapter 3. What is being detected by the GEMS?

In this section I will relate some of the objectives of the experiments described in this chapter as well as describe the various hypotheses the experiments were designed to test. The experiments will be described in four sections: The procedure, objectives, analysis, and conclusions.

The main objective of the experiments was to determine what the voltage return from the GEMS represented in terms of physiological motion, i.e. what tissue motion was causing the GEMS signal. To accomplish this, we must understand how the GEMS senses changes in position (see Section 2.1.1), how speech is produced (Section 2.3 and 2.4), and how electromagnetic radiation interacts with dielectric substances (Section 3.2). Only then are we in a position to understand how the GEMS and the tissue interact.

It is also important to understand how the GEMS has been used in these experiments. Figure 3.1 demonstrates the vertical position of the GEMS along with possible locations for separate EM sensors that have been used to measure the motion of the jaw, lips, and tongue (Holzrichter et al. 1998).

Section 3.1. Theories proposed

We will now inspect three competing theories on the physiological basis of the return that arose before and during the various experiments.

Reflection from fold surfaces

This was the first theory, proposed by the researchers to first examine the GEMS and its signal. It is perhaps the most obvious – the folds are moving so they must be causing the signal. In this hypothesis, the GEMS return would be proportional to one or

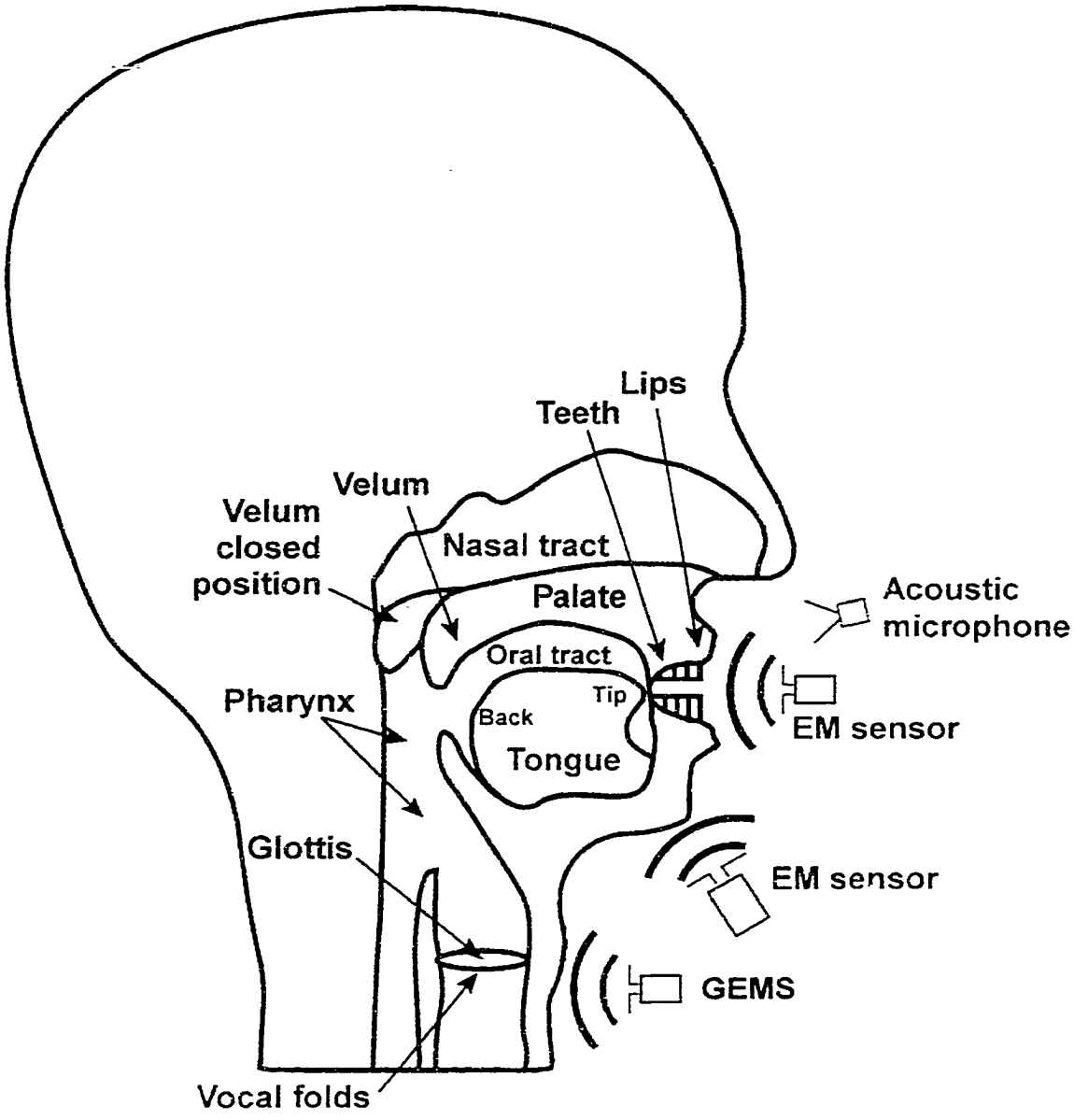


Figure 3.1. Use of the GEMS and other EM sensors to detect human vocal articulator movement.

both of the folds' motion. The area of the glottis and the resultant airflow would have to be derived from the fold position vs. time.

Transmission through folds

This theory was formed after the GEMS was compared to the electroglottograph (EGG), during experiments performed at the University of Iowa's National Center for Voice and Speech. The experiments were performed in collaboration with Drs. Ingo

Titze, Brad Story, and Wayne Lea. They will be discussed in detail in Section 3.4. The EGG is an instrument that causes radio frequency (MHz) AC current to flow from one electrode to another. The electrodes are not co-located (i.e. the bistatic configuration is used, unlike the GEMS antennae). An electrode is placed on either side of the trachea by the laryngeal prominence. The current passes through the folds when they are in contact and is greatly attenuated when they are not. The resulting signal is proportional to the amount of vocal fold contact area. Upon comparison of the EGG signal and the GEMS signal, many similarities were noted. It was upon this basis that the theory was formed that the GEMS might be acting in a similar manner – transmitting the microwave (GHz) signal into one fold and out the other. In this manner the GEMS might be considered to be a portable, inexpensive EGG.

Reflection off of the tracheal wall

This was the theory I proposed in the summer of 1995. In this hypothesis the great majority of the GEMS energy reflects off of the anterior and posterior walls of the trachea, resulting in a signal that is proportional to trachea vibration and thus subglottal pressure. This theory arose from countless hours of experimenting with the GEMS in a variety of situations with a variety of targets: 2 cm diameter mirrors, speaker cones, and various other vibrating objects. These experiments were done at the very beginning of the project in an attempt to understand how the GEMS reacted to vibrating objects. In all of them the GEMS exhibited large responses for very small (less than a millimeter) amplitude vibrations (in air) less that a centimeter away from the antennae. This (and some simple E&M calculations, see Section 3.2) led to the conclusion that the vibrations of the trachea would be sufficient to generate the signal produced by the GEMS.

More recently (see Section 2.1.3), we have seen that the GEMS can detect motion on the order of micrometers at distances of centimeters, so it is certainly possible that the GEMS is detecting the vibration of the tracheal wall.

Once the relationship between the GEMS signal and the trachea-wall-driving subglottal pressure is established, the subglottal pressure can be derived using the GEMS information. The subglottal pressure can then be used to define an excitation function of the vocal tract. With an accurate excitation function and good audio output, the human vocal tract transfer function can be described reliably and with good accuracy. This is explored in Chapter 4.

Section 3.2. Electromagnetic calculations and simulations

In this section we will discuss the E&M techniques used (both simple and sophisticated) to model the behavior of the electromagnetic waves as they interact with the complex tissues of the human body.

3.2.1. Dielectric properties of human tissue

The first thing that must be done is to get an idea of the dielectric properties of human tissue so that a meaningful model can be constructed. Several sources were consulted, but most of the studies were done *in vitro* (out of the body) after death. There was very little data (for good reasons) on live human tissue, specifically skin, fat, cartilage, and smooth and skeletal muscle. These are the principle components of the tracheal region (see Section 2.2). Most of the values used in research come from canine or porcine (pig) tissue.

The values we need in order to run a simulation are the permittivity ε, conductivity σ, and permeability μ of the tissue. The permittivity describes how the material is affected by electric fields and is usually described as a dimensionless ratio in terms of ε_o, the permittivity of free space. This ratio of $\varepsilon/\varepsilon_0$ is termed the dielectric constant, an unfortunate name as it can vary substantially with frequency and can therefore only be considered constant for a finite range of frequencies. Thus, a dielectric constant of 52 would mean the permittivity of the tissue is 52 times that of free space at the frequency of interest. Since the displacement field D is equal to εE (see Griffiths (1989) for this and other details concerning dielectric properties) and is continuous along dielectric boundaries, the strength of the electric field E inside the tissue would be 52 times less than the field outside if all electromagnetic energy transferred to the tissue. This effect is

due to polarization of the tissue, which opposes the electric field inside of a dielectric. If we represent the electric field in free space with a E_0, and the electric field in the dielectric with E_1, then at the boundary of the dielectric

$$D_0 = D_1$$

or

$$E_0\varepsilon_0 = E_1\varepsilon_1$$

$$E_1 = E_0 \cdot \frac{\varepsilon_0}{\varepsilon_1}$$

Since the permittivity of the dielectric is always greater than that of free space, the electric field is decreased by the dielectric constant.

This process is similar to the polarization of magnetic dipoles that occurs in magnetism. This is in fact what permeability is – the magnetic analog of permittivity. Since human tissue is normally relatively free of magnetic materials, the permeability for all simulations was set to the permeability of free space.

Conductivity is another matter altogether. Conductivity is the inverse of resistance and describes how much the electromagnetic wave is attenuated as it transits the material. Conductivity destroys electromagnetic radiation as the free electrons and ions in the material are moved by the incoming wave's fields. The electrons and ions will seek the lowest energy state and rearrange themselves in such a way as to cancel out the incoming wave's field. Whether they are successful or not depends on how good the conductor is and whether resistance is present. The better the conductor, the faster and more complete this cancellation. In a "perfect" conductor, the electrons are infinitely fast and electromagnetic waves cannot penetrate at all. If the electrons and ions can lose energy

through collisions, the electromagnetic energy (that has been converted into kinetic energy) is dissipated by the collisions and turned into heat. This is the process by which a microwave oven heats food, through the excitation of water molecules.

An example of how conductivity attenuates electromagnetic waves is the transmission of light into seawater. Seawater is not a good conductor (about a million times less conductive than most metals), but light is unable to penetrate more than 30-40 meters. It is not due to sediments or suspended material (although that does make it worse), the electromagnetic waves are simply dissipated by the free electrons and ions in the water. In the process the water is heated. Keep in mind that conductivity is a function of frequency as well, but again we will be idealizing our waves to a single frequency so for our purposes the conductivity is constant. The conductivity of human tissue in our frequency range of 1-3 GHz is about 1.1 to 2.5 S/m (Duck 1990, Table 6.13; Lin (1986)), about 10-20 times less than sea water and about even with germanium, a semiconductor (Griffiths (1989), p. 271).

In order to choose valid approximations to the dielectric "constants" I used values taken again from Duck (1990), Lin (1986), and Haddad *et al.* (1997). The values are not really constant as they change with frequency, but we will use the average values from 1-3 GHz and assume they are relatively constant over that region of interest. Agreement between sources is fairly good, with the difference in values attributable mostly to the moisture content and whether or not the measurement was in-vitro (out of the body) or in-vivo (in the body) and whether the measurement was taken before or after death. Canine and porcine tissues are quite similar to human, and were used as an approximation in some cases so that before-death in-vivo measurements could be taken.

Freq in GHz	λ_o cm	Fat			Cartilage			Smooth muscle			Skin		
		ε_r	σ	d	ε_r	σ	d	ε_r	σ	d	ε_r	σ	d
1	30	15	0.2	10.4	22.5	0.12	21.0	56	1.4	2.8	33.5	0.7	4.5
2.5	12	13	0.4	10.3	21	0.13	21.0	52	2.2	1.8	33	1.0	3.1

Table 3.1. Dielectric constant and conductivity for biological tissues at approximately 1 and 2.5 GHz (Duck (1990), Lin (1986), Haddad *et al.* (1997)). ε_r is relative dielectric constant, σ is conductivity in S/m, and d is the skin depth in cm.

3.2.2. Plane-wave scattering from planar surfaces

This simple calculation was first performed in 1995 by the author in an attempt to understand how the wave was reflected from the different dielectric layers in the glottal area. For simplicity, conductivity and geometrical factors were ignored. That is, the wave is assumed not to be attenuated as it travels through the dielectric materials and all layers are considered planar with arbitrary thickness. This will lead to reflection coefficients that are too high, as the layers are somewhat conductive and some shapes (like the folds) are not good reflectors in the direction of propagation (from the anterior to the posterior of the neck). With conductivity neglected, the dielectric constant of interest is the permittivity, or in this case the *index of refraction* n:

$$n = \sqrt{\frac{\varepsilon\mu}{\varepsilon_o\mu_o}}$$

or, if we use ε_r (the ratio of tissue permittivity to free-space permittivity) and assume a permeability of $\mu = \mu_0$:

$$n = \sqrt{\varepsilon_r}$$

The percentage of energy reflected at each planar surface (the reflection coefficient R) is determined by the change in the index of refraction by the following formula (Griffiths, 1989):

$$R = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2$$

where the wave is travelling from material 1 to material 2. The percentage of energy transmitted at each surface (transmission coefficient T) is simply $1 - R$.

Figure 3.2 shows the setup of the simulation and the results. The neck is represented as several layers of materials with different indices of refraction. The smooth muscle represents the trachea or the folds, both being of similar materials. There is an air gap to represent either the tracheal air or the glottis, and a final layer of smooth muscle for the far tracheal wall or second fold. Interference effects between layers are neglected as a second-order effect, so the width of each layer is arbitrary. At each interface the reflection coefficient is labeled.

It is evident that most of the energy (54%) is reflected from the outer surface, but normally the GEMS's plastic case is held lightly against the skin, so this reflection is considerably reduced. In addition, there is evidence that most of the signal is due to tissue motion under the skin. I conducted an experiment (discussed in more detail in Section 3.5) where a copper foil was placed on the outside of the skin to magnify any return originating there. The foil reflects nearly 100% of the incoming EM energy (being a very good conductor), blocking the EM energy from the interior of the throat. The GEMS was held about 1 cm from the skin surface, and the amplitude of the GEMS signal was noted with and without the foil. The amplitude of the GEMS signal dropped by a factor of 8 with the foil in place, indicating that the majority of the signal originated form inside the neck.

So what structure inside the neck would be most likely to reflect electromagnetic energy? The fatty, cartilage, and smooth muscle are not terribly different in n, so R is not

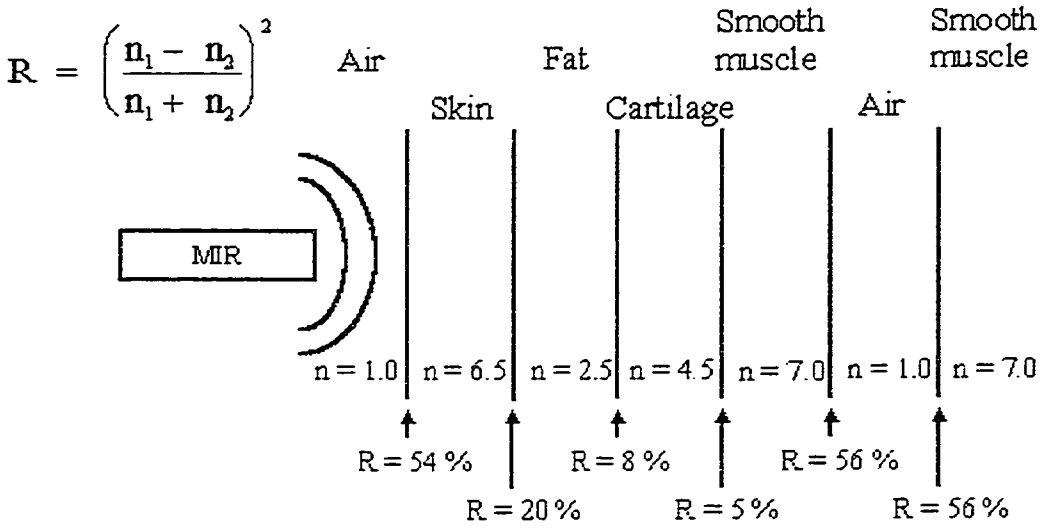


Figure 3.2. Simple planar calculation of the neck tissue layers' reflectivity, neglecting geometrical factors, multiple reflections, and conductivity.

large at those interfaces. The large values of R come from where tissue abuts air, at the tracheal walls or fold surface. Therefore, we may conclude that most of the energy comes from one of those two regions.

If we take geometrical considerations into account, the tracheal wall seems much more likely to reflect more energy than the trachea. The walls (especially the posterior one, see Section 2.2) are large and relatively flat compared to the wavelength of the transmitted wave. The folds, however, even in their most open position, are shaped much like the hull of a boat. The waves are incident upon the "bow", where the frontal area is quite small. The change in area "seen" by the wave would be quite small. The folds are simply not shaped to reflect anterior-to-posterior travelling waves with any sort of efficiency (see Figure 3.3).

The reduction of the wave due to the conductivity of the tissues can be approximated by the calculation of the skin depth. The skin depth can be thought of as a metric of the amount of energy loss that electromagnetic waves suffer in conductive

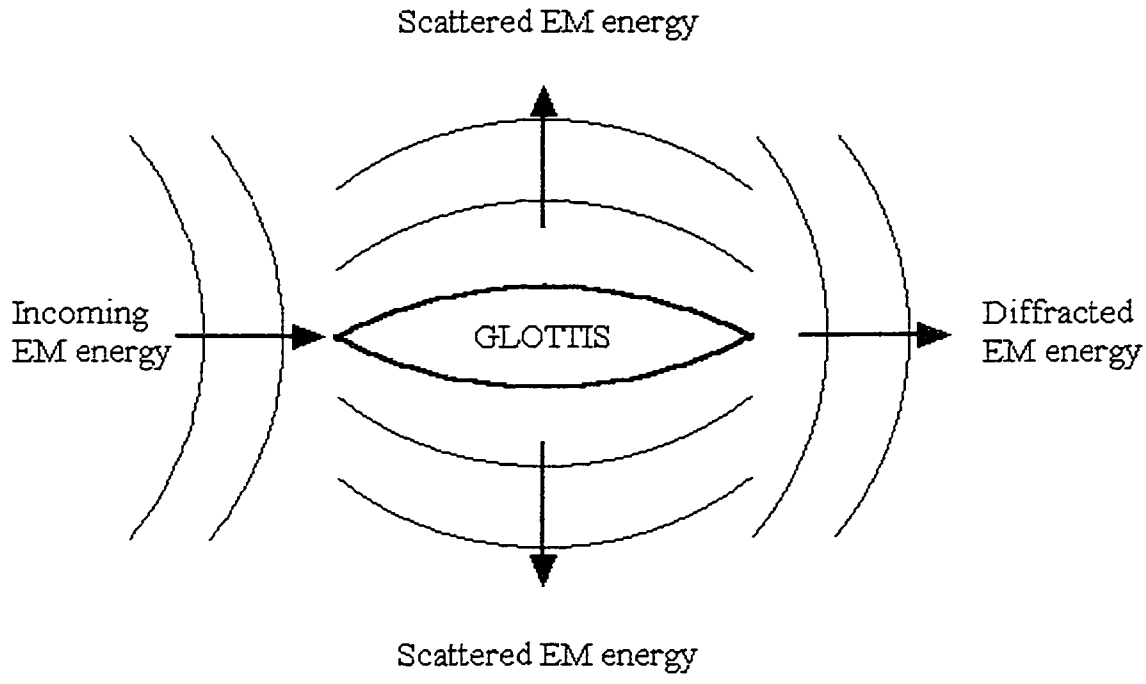


Figure 3.3. Demonstration of the geometrical effects on EM wave scattering from the folds. As viewed front the anterior side, the folds have little scattering cross-section. Most of the energy is simply diffracted around the folds. Where scattering does occur it is not reflected back to the transmitting antenna but rather to the sides.

materials. It is the depth in the material at which the amplitude of the electromagnetic wave has been reduced by 1/e. The formula given by Griffiths (1989, p. 370) for skin depth d is:

$$d = \frac{1}{k_-}$$

where

$$k_- = \omega\sqrt{\frac{\varepsilon\mu}{2}}\left[\sqrt{1+\left(\frac{\sigma}{\varepsilon\omega}\right)^2}-1\right]^{1/2}$$

Using the values for smooth muscle at 2.5 GHz given above, the approximate skin depth is 1.8 cm. This means that for every 1.8 cm in tissue the wave has to travel, it loses about 63% of its energy to the conductivity of the tissue. The skin depth is greater for

other tissues (such as fat and cartilage) due to their lower conductivity. Calculated skin depths for various tissues at 1 and 2.5 GHz are given in Table 3.1. It is an important distance to keep in mind as we look at the following simulations. As far as the trachea is concerned, it is not a major impediment as the tracheal wall is only a few millimeters thick. However, the skin, fat, and neck muscle between the antennae and the trachea are on the order of 1 cm thick, so significant attenuation of the wave occurs. Also, each fold is about 1-2 cm long, so if the wave had to traverse part of the folds, reflect off a surface, and go back through the folds, the skin depth could be traveled quite easily. The skin depth becomes an important quantity for experiments conducted where the GEMS was pulled off-center around the side of the neck (as will be seen in section 3.4), as many more centimeters of tissue must be traversed before the trachea or folds can be reached. Due to skin depth conduction losses, a significant attenuation of the GEMS strength was observed.

3.2.3. 2-D finite-element electromagnetic simulation using TSARLITE

Procedure: These experiments were designed to calculate the relative reflectivities of the trachea and vocal folds using Jeff Kallman's excellent 2D E&M simulation program, TSARLITE. TSARLITE is a 2-D rectangular-cell finite-difference time-domain modeling code. It directly solves Maxwell's equations in two dimensions. The tissue was specified by a dielectric constant of 52, a conductivity of 1.1 S/m (average values for muscle from Table 3.1), and a permeability equal to that of free space. The values (except for permeability) were varied slightly to observe the effects. The waves were specified by wavelength, duration, and envelope (usually Guassian). Simulations were normally done at 2.3 GHz (at the time believed to be the center frequency of the

GEMS, more recent tests have shown it is closer to 2.1 GHz), but some were calculated at 915 MHz, a frequency under consideration for future versions of the GEMS.

The simulations discussed here were done in a continuous dielectric material with an air-filled hole shaped like the trachea or the folds as the scatterer. This was done to clarify the differences in the folds' and the trachea's ability to scatter EM energy back to the transmitting antenna. Other interfaces (such as the skin and outside air) were modeled in earlier experiments but multiple reflections cluttered the results and made it difficult to determine where majority of scattering was taking place.

For each configuration, a "movie" consisting of 40 to 100 frames was generated with colors representing electric field strength. This is useful for giving a feel of how the electromagnetic fields propagate through the model, but does not generate quantitative results.

To that end, a vertical line was established on the grid where the electric and magnetic fields were recorded at every point on the line at each time step. This allowed the calculation of Poynting vectors and subsequently reflection coefficients. However, the generation and analysis of such lines was computationally intensive and was therefore only done on four occasions. In the reflectivity calculations, for simplicity, the entire grid was filled with tissue and the trachea and folds were simulated with air holes. The trachea was represented by a circle 1.5 cm in diameter (with a flattened rear wall), and the folds by an oval with a maximum width of 6 mm, far wider than would be expected for normal speech. The waves were launched from the left side and propagated to the right. Figure 3.4 illustrates the simulation configuration. It consists of a single frame from a tracheal experiment movie.

Objectives: The primary objective was to get an idea of the difference in reflectivities for the simulated trachea and folds. A secondary objective was to observe the role diffraction played in the interaction between the wave and the reflectors.

Movie Analysis: Each movie was read into a Sun workstation (by my colleague Melinda Bass) using XTRAS, a program supplied by Jeff Kallman. As each frame was displayed, it was window dumped and saved as an .xwd file. The frame files were then transferred via ftp to a computer where they were reassembled into movies playable on a PC inside Matlab 5.21. There were 12 movies recorded in total.

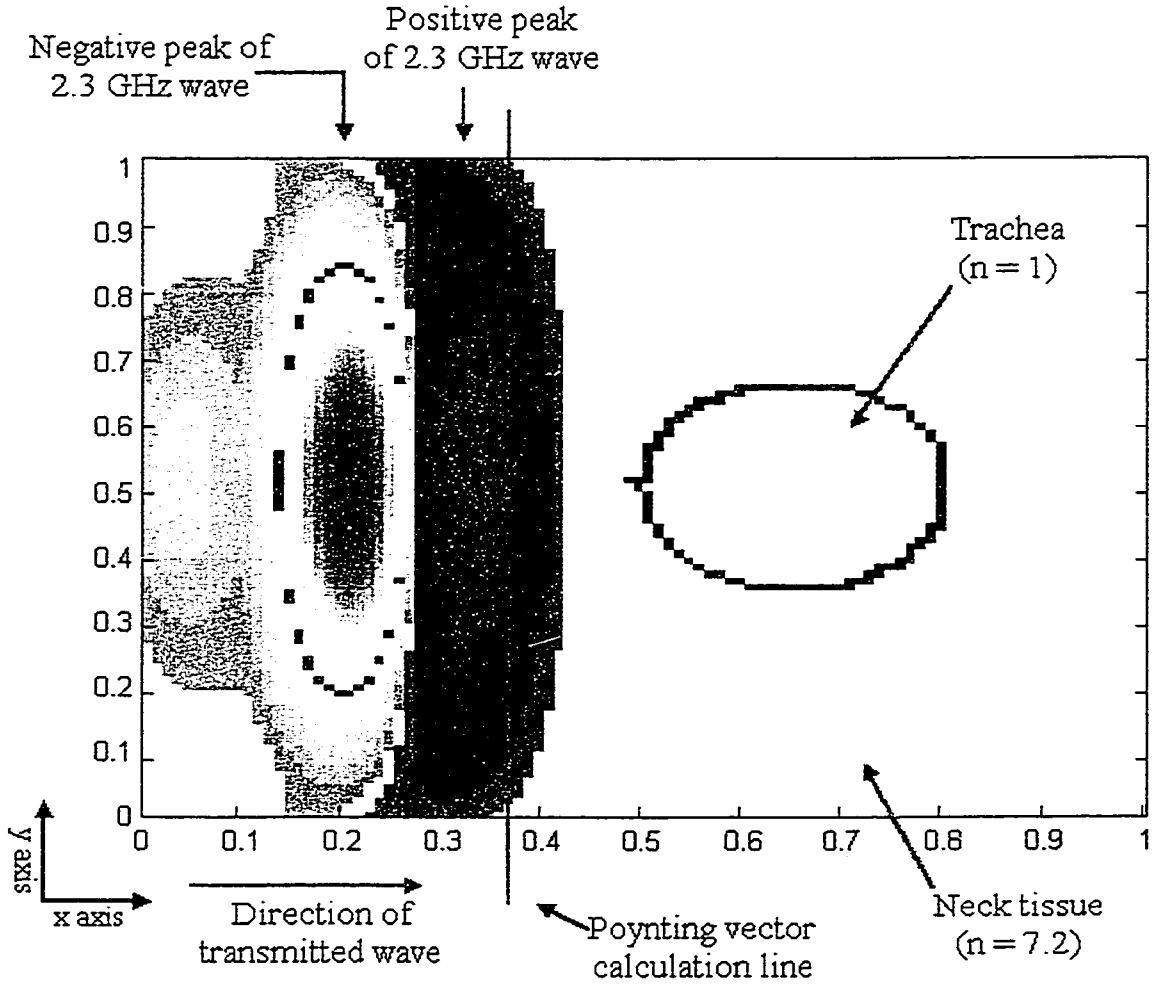


Figure 3.4. Frame from tracheal reflectivity simulation with 2.3 GHz wave. The frame is slightly stretched in the x direction due to machine graphics incompatibilities.

Results:

1) All of the movies exhibit a good deal of diffraction around the trachea and folds, but the waves diffract much more easily around the folds than the trachea. This is due to the smaller frontal cross-sectional area of the folds as seen from the anterior portion of the neck discussed earlier. The diffraction was reduced somewhat by the reduction of the wavelength of the GEMS signal in the dielectric by $\sqrt{\varepsilon} = \sqrt{52} = 7.2$ from about 13 cm to 1.8 cm. This is about the same as the diameter of the trachea, but three times the maximum width of the folds. As a result, the wave front observed in the neck after transiting the folds was considerably more uniform than the wave front observed after passing by the trachea. The transmitted tracheal wave front had a significant gap in it, presumably where the GEMS energy had been removed by reflection off the anterior and posterior surfaces.
2) The diffractive effect was much stronger in the longer wavelength 915 MHz simulations. This is expected due to the increase in wavelength. This would indicate that the 915 MHz waves will not be as efficient as the 2.3 GHz waves in detecting motion of objects on the order of 1-2 cm in size. However, the longer wavelength will more than double the distance between the null points of the GEMS sensitivity envelope (see Section 2.1.3), where it changes sign. This should result in a more stable signal.
3) In one simulation, the folds were positioned 45 degrees from the left fold to better simulate any forward transmission through the folds that occur, given favorable circumstances. This was to test hypothesis 2, that the GEMS EM wave is

transmitted through the folds similar to an EGG signal. It was clear that most of the energy diffracted around the folds, with the rest reflecting off the left fold/air boundary. Only a minute amount (the lowest value possible except for background) ended up traveling away at 45 degrees from the right fold as if it had been transmitted through the folds. It is just not geometrically possible at this wavelength for the folds to "pull" the energy in, transmit it across the glottis, and "push" it back out the other side. The index of refraction is not high enough for this to occur. This argues against the transmission theory discussed in the introduction to this chapter.

Reflectivity results

For quantitative reflectivity measurements, Dr. Kallman's .empp file was used to record all relevant EM components (in this case H_x, H_y, and E_z) along a vertical line for every time step. The resulting file contains a header describing the data and all the data points in ASCII format. After splitting the .empp files into a header and a data .txt file, it is possible to read the data into Matlab and analyze the results.

Knowing the **E** and **B** fields at every point, I could calculate the Poynting vector **S** by taking the cross product of the two:

$$\vec{S} = \frac{1}{\mu_0}\left(\vec{E} \times \vec{B}\right)$$

The Poynting vector is the energy flux density – it tells us how much and in what direction the energy is traveling per unit time per unit area. To get the total energy per unit time (or power) traveling through the vertical line, we integrate along the line:

$$\frac{\text{Energy}}{\text{time}} = \int \vec{S} \cdot dl$$

or, since we are in the discrete domain, the power passing through the line at each time point is the sum of the individual $\hat{x}$ components of **S**. We use the $\hat{x}$ component since the TSARLITE grid is set up in Cartesian coordinates, with the vertical line composed of constant x. At each time step we sum the S_x components along the line. That tells us the amount and direction of the energy flux through the line at that time. To get the total energy passing through the line, we sum the energy flux over time. For a perfect reflector, the total energy would be zero (the incident energy would equal the reflected, or equivalently there will be equal amounts of negative and positive S_x) and the coefficient of reflection R would be unity. For anything else, there will be some energy transmitted through the medium and R will be less than one. To calculate R we simply determine the ratio of reflected ($-\hat{x}$ direction) energy E_{out} to incident ($+\hat{x}$ direction) energy E_{in}:

$$R = \frac{E_{out}}{E_{in}} = \frac{\sum_t \text{negative} S_x}{\sum_t \text{positive} S_x}$$

This formula assumes that no energy is lost in the incident or reflected waves in the $\hat{y}$ direction as they travel between the line and the trachea/folds. This is naturally not completely true, but as the pulse is originally traveling in the $+\hat{x}$ direction and the line is somewhat close to the trachea/folds (12.5 mm away) we can be assured that most of the relevant energy will pass through the line on its way in and out. In these tests the conductivity was set to zero, so any reflection will be due to dielectric differences alone.

Results

The .empp method was used for the following situations:

1) Solid block of tissue, frequency = 2.3 GHz (calibration run)

Experiment #	Reflected E	Incident E	R (%)
1: Calibration block	40.7	70.5	57.6
2: trachea at 2.3 GHz	6.1	39.8	15.2
3: folds at 2.3 GHz	0.3	39.8	0.8
4: trachea at 915 MHz	4.9	195.7	2.5

Table 3.2. Reflectivities of various configurations modeled in TSARLITE.

2) Trachea 1.5 cm wide, frequency = 2.3 GHz

3) Folds 15 mm long, 6 mm wide, frequency = 2.3 GHz.

4) Trachea 1.5 cm wide, frequency = 915 MHz

The results are summarized in the Table 3.2, with the graphs of the S_x versus time plots given for experiments 1, 2, and 3 at the end of this section.

The calibration test was done with a single slab of $\varepsilon = 52$ tissue on the left and air on the right. The amount of energy reflected from a perpendicularly incident wave is:

$$R = \left(\frac{n_2 - n_1}{n_2 + n_1}\right)^2 = \left(\frac{-6.2}{8.2}\right)^2 = 57.2\%$$

which measures up very well with the 57.7% calculated. Thus we are ensured that our findings for the other models are similarly accurate. The results are shown in Figure 3.5.

The energies given are in Joules $\cdot \mu_0$, as the results were not divided by μ_o in order to make the numbers more readable. There are two situations that facilitate direct comparison: number 2 (Figure 3.6), and number 3 (Figure 3.7). They were both done at 2.3 GHz and the reflectors (the trachea in number 2 and the folds in number 3) are the same diameter, just different shapes. The difference in R is striking – the trachea reflects

back to the line almost 19 times the energy of the folds. This indicates that the folds are not the best reflectors of EM energy in the glottal area. Indeed, in this simulation the conductivity of the folds is neglected, and the conductivity of the muscle tissue is known to be relatively high ($\sigma \approx 2.2 \, S/m$, skin depth of about 1.8 cm). Thus to traverse through the folds, reflect off the medial surface, and travel back to the receiver (even if geometrically possible) would reduce the energy by a further 65% or so. In addition, in the "movies" of frames made from this simulation what little energy does reflect does so at the front of the folds, with almost no energy reflecting from the sides of the folds. Most of the wave simply diffracts around the folds.

In addition, when 3D effects are considered, the folds become even less likely to contribute to the GEMS signal. They are only on the order of 1 cm from top to bottom, and as such do not present a large reflective area. The trachea, on the other hand, extends for several centimeters down into the chest (and vibrations may be detected reliably all the way down the trachea) and a reflective surface exists for many centimeters above the glottis as well.

Experiment 4 demonstrates the effect diffraction has upon the reflectivity. For the same sized object (the trachea), the reflectivity drops from 15.2% to 2.5% - a reduction of sixfold from an increase in the wavelength of just under 2.5 times. This indicates that operating at 915 MHz may result in a reduced signal magnitude due to increased diffractive effects.

Conclusions

The first conclusion that can be drawn is that at the wavelengths of interest, the EM waves have little interest in reflecting from either the trachea or the folds. The majority

of the energy of the waves simply diffracts around the objects, even though the objects have about the same dimensions as the waves. The evidence from this comes from the qualitative evaluation of the movies of different configurations as well as the comparison of scattering efficiency vs. the frequency of incoming wave. For the trachea at 915 MHz, R was six times less than that measured at 2.3 GHz. The longer wavelength at 915 MHz diffracted even more readily around the trachea.

Comparing the two structures (trachea vs. folds) directly, the energy measurements show conclusively that the tracheal wall is far more efficient at reflecting energy, returning 19 times the energy back to a spot 12.5 mm in front of the reflector. This clearly indicates that the trachea is more likely to be causing a majority of the signal.

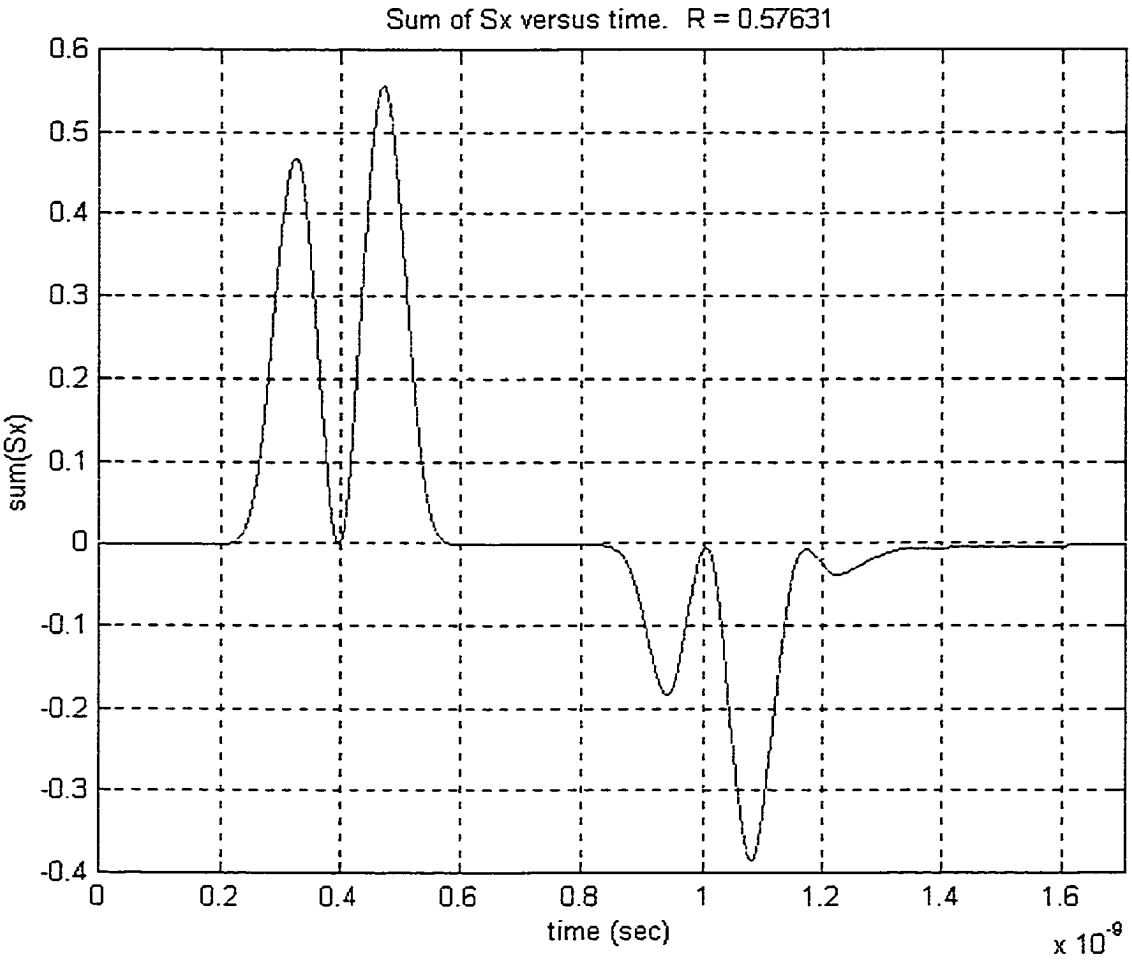


Figure 3.5. Energy vs. time for the calibration experiment. Positive values are energy moving to the right (incident), negative values are energy moving to the left (reflected). The two peaks are the positive and negative fields peaks shown in Figure 3.4. In this example, R = 57.7%, very close to the theoretical value of 57.2%.

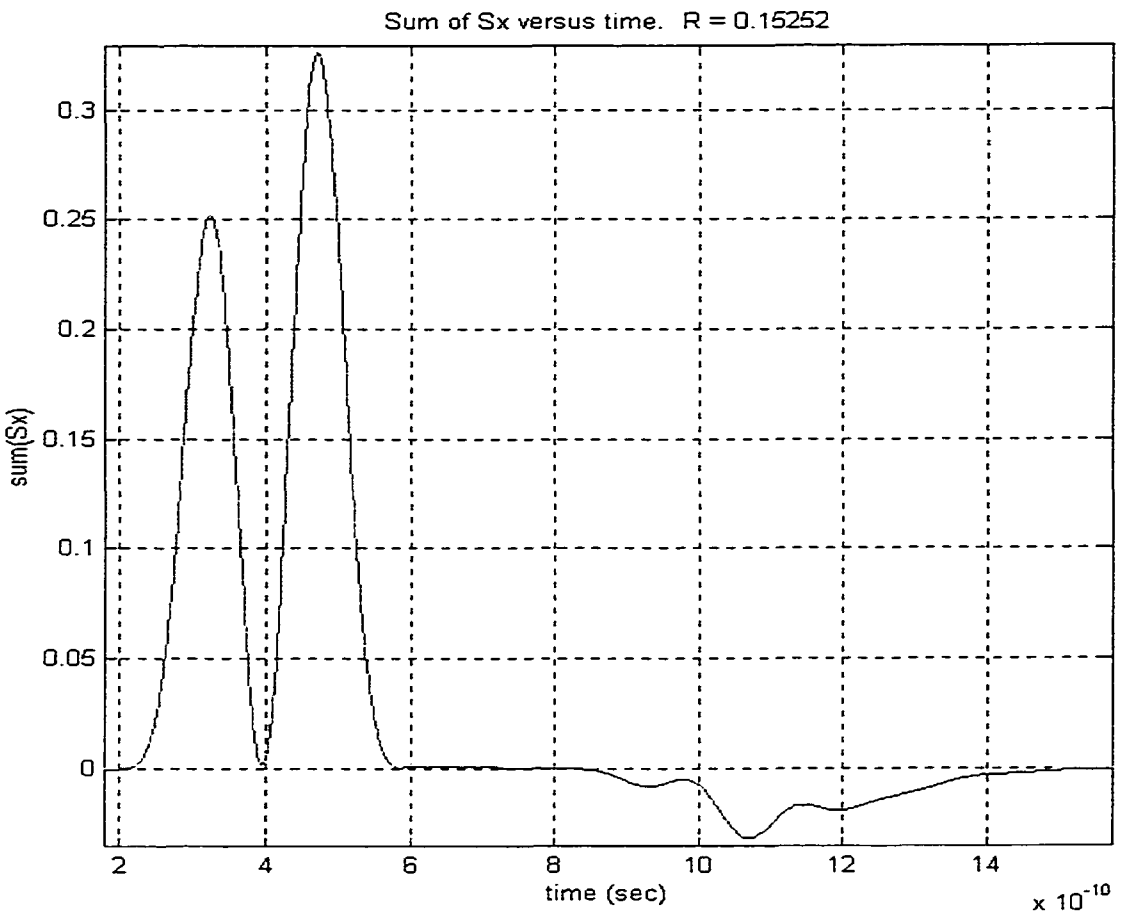


Figure 3.6. Energy vs. time for the trachea experiment. Note the reflected energy from the trachea between 0.9 and 1.4 nsec. R = 15.3%.

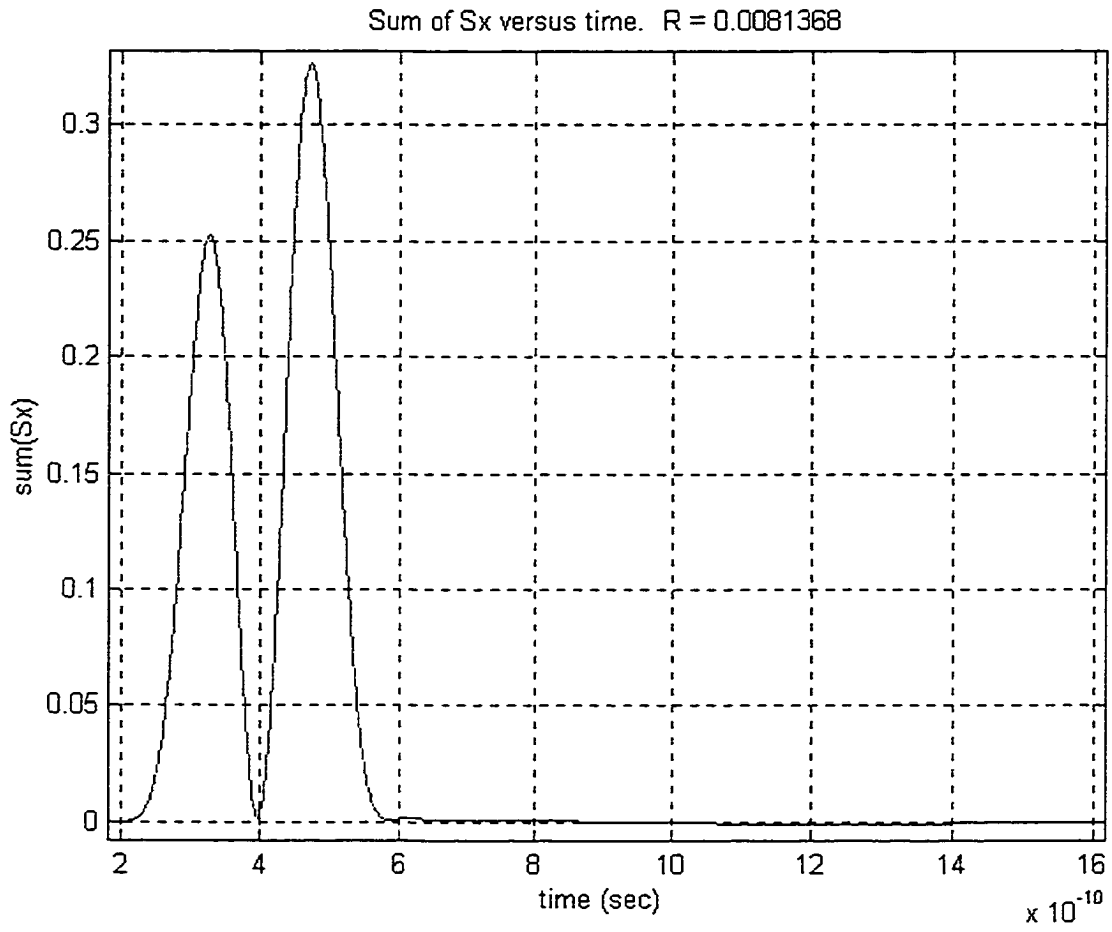


Figure 3.7. Energy vs. time for the fully open folds experiment. Note the almost total lack of (negative) reflected energy. R = 0.8%.

Section 3.3. High speed video experiments

Procedure: On three separate occasions, high-speed video experiments were conducted at the Department of Otolaryngology at the UC Davis Medical Center with Dr. Rebecca Leonard. A Kodak EktaPro motion analyzer model 1012 with intensifier was used with a laryngoscope to capture high-speed (1000 and 3000 fps) videos of the vocal folds in motion, while simultaneously recording audio and GEMS signals. The data was again recorded through Labview 4.0 and the same National Instruments A/D as described in Section 2.1.3 was used. The sampling rate was normally 40 kHz, although 10 kHz was used on some of the earlier experiments. If taken at 40 kHz, the data (except the frame marker, which required a high frame rate as its pulse was only 30 microseconds (μsec) wide) was digitally filtered and decimated to 10 kHz for analysis. A frame marker denoting the exposure time of each frame was generated by the camera and upon analysis was superimposed upon the GEMS signal to facilitate comparison between the corresponding video frame and the GEMS signal. A segment of GEMS signal and some of the corresponding video frames is shown in Figure 3.8. For more information on the experimental setup, please see Appendix B, where more information (and a paper now under review for the journal "Phonoscope") has been added.

Objectives: The primary objective of these experiments was to see how the fold motion related to the GEMS-derived position signal for a normal adult speaker under various speaking conditions and registers. A secondary objective was the comparison of the GEMS signal of a patient with prominent vocal fold nodules to the videos. Both tests were meant to determine how the folds influenced the GEMS signal. The questions to be answered were as follows:

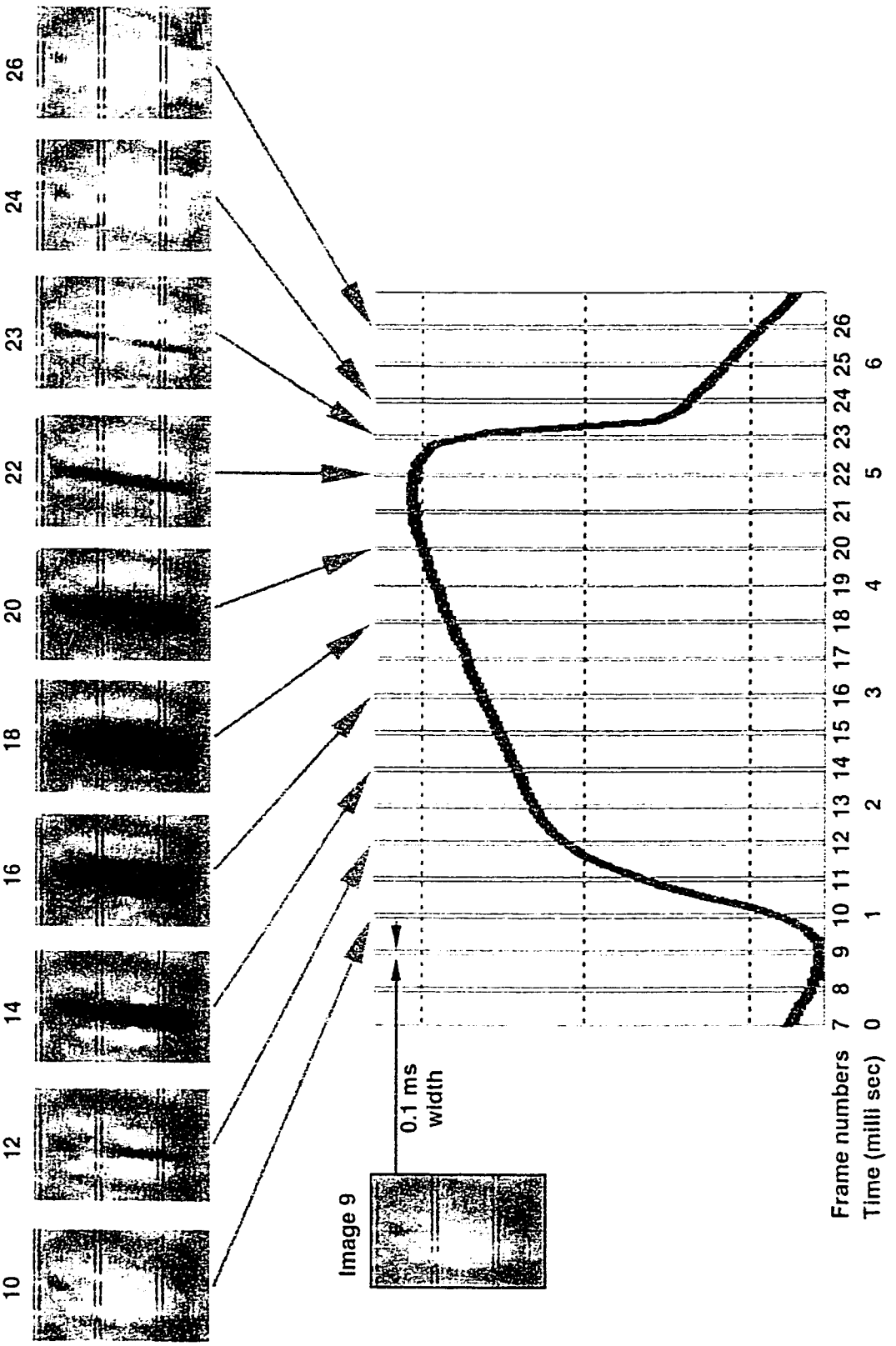


Figure 3.8. One cycle of a GEMS signal with some of the corresponding video frames. The vertical bars superimposed on the GEMS signal denote the exposure time. The GEMS signal has not been inverse filtered. The horizontal bars on the video frames are caused by a camera defect.

1) Does the fold motion correlate with the GEMS return, or does the GEMS seem to match what might be sub- or supraglottal pressure variations?
2) Can we detect a difference in the GEMS return due to the presence of a large nodule on one fold? As one nodule is larger than the other, if we can detect their presence, how will the asymmetric distribution of tissue affect the return?

Analysis of abnormal physiology: We will begin with the analysis of the patient with the vocal fold nodule, as the results were immediately apparent. The patient had a nodule on his right fold, which had in turn formed a smaller nodule on the left fold through friction. The effect on his voice was to make it quite raspy, with occasional very brief lapses of almost no sound at all. Throughout the experiments, the subject's audio alternated between "poor" and "normal" sounding audio cycles. The poor cycles produced very little audio energy at all. As a result, the combined poor and normal cycles sound like a series of pulses, rather than a steady tone. This pulsation reduces the perceived pitch of the signal by a factor of two and results in a raspy voice.

The subject attempted to hold a steady tone of about 130 Hz. After analyzing several thousand frames of video with the respective data, I have made many observations, which I will group by their location on the inverse filtered GEMS (which represents reflector position) waveform. See figures 3.9 and 3.10 for those locations and a quick summary of the observations made.

Top of GEMS return and rapid drop (~frames 29-31 in Figure 3.9)

At the beginning of this region (see Figure 3.10 regions 2-4, also Figure 3.9 at about frame 29) the middle nodule opens and separates from the left fold. There is no discernable change in the GEMS, its derivatives, or its integral to indicate the nodule has

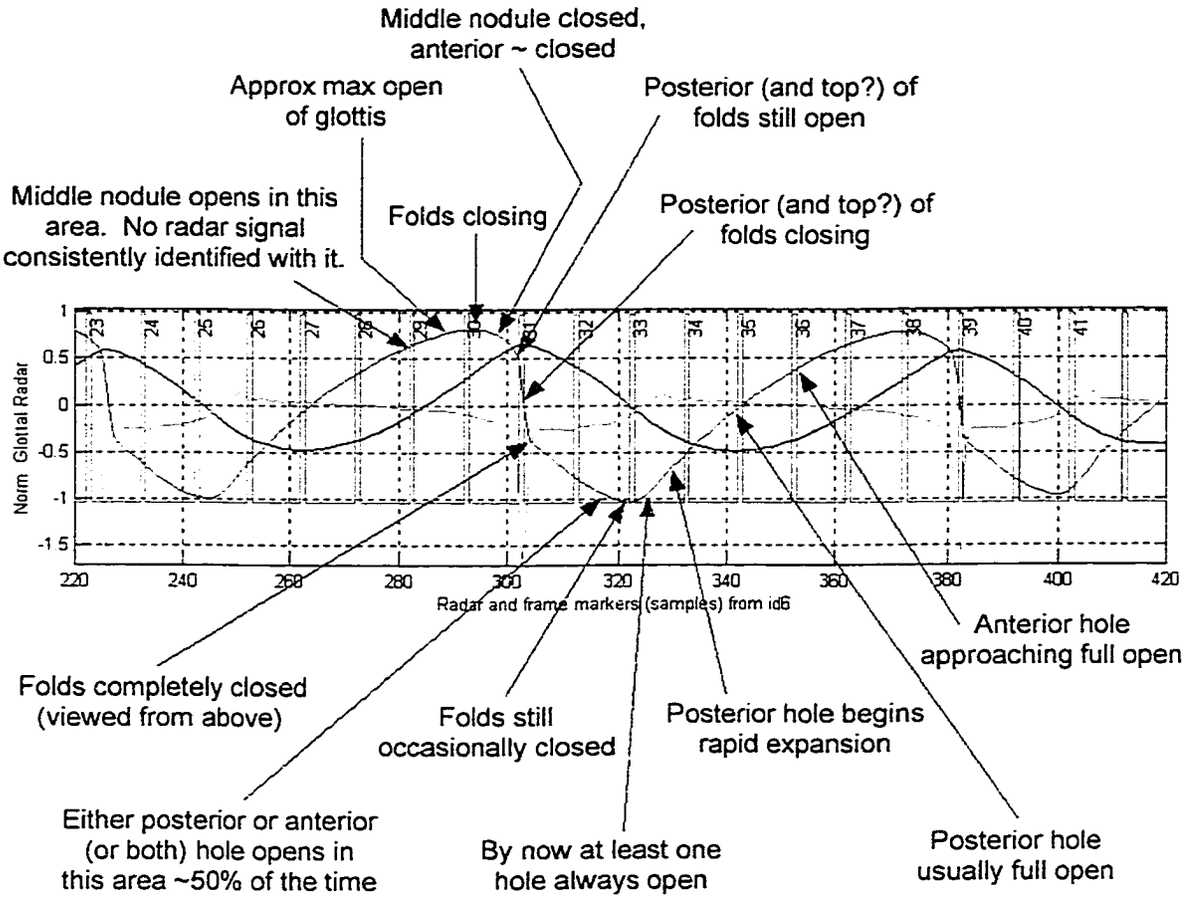


Figure 3.9. Plot of inverse filtered GEMS (blue), first derivative of GEMS (green), and integral of GEMS (black) along with the frame markers (red) for the abnormal physiology subject. The width of the frame markers denotes the exposure time of the frame. Observations for the dataset are included. The time scale is in samples, at 10000 samples/second.

broken contact with the left fold. This observation was confirmed in several of the other videos. This implies that vocal fold contact is not registered by the GEMS. As contact should cause a large disruption in the GEMS signal if transmission or reflection from the folds is occurring, this is strong evidence that the folds do not directly affect the GEMS signal.

Closer to and at the maximum of the GEMS return (Figure 3.10 regions 5-7), I found that the maximum of the cross-sectional area of the folds actually occurs about 0.3-0.5 milliseconds before the maximum of the GEMS return. This result was confirmed in

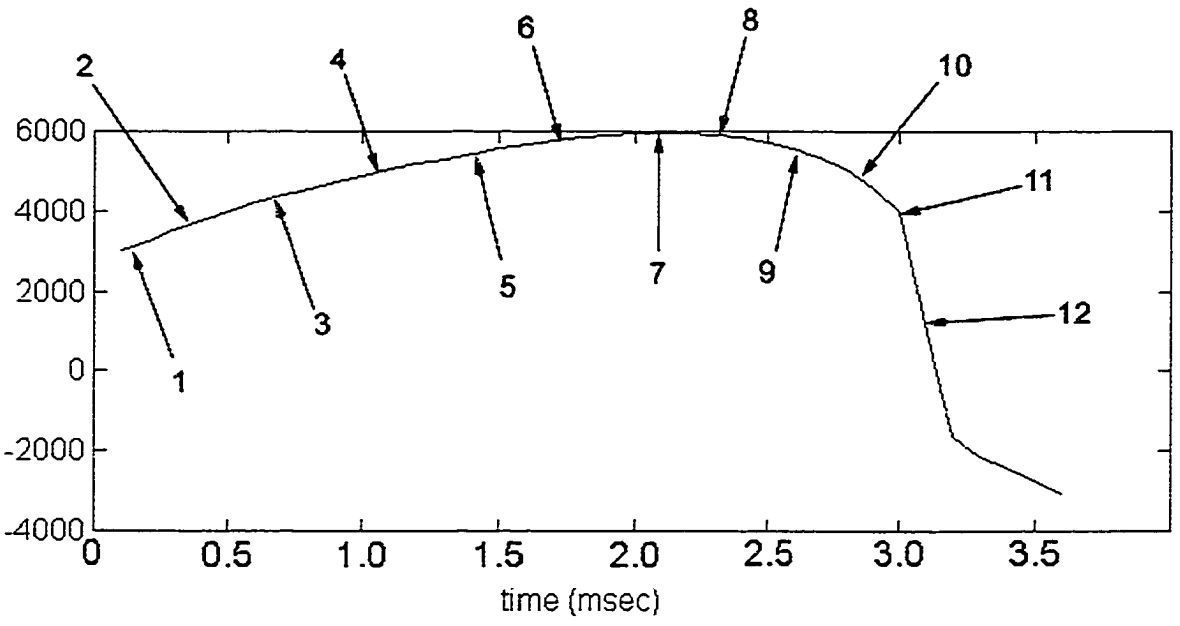


Figure 3.10. Approximate locations on the GEMS return for the frames analysis.

studies of normal folds that will be discussed in the next section. This is an important observation, as it means the folds have stopped their motion and reversed course without any indication from the GEMS that they have done so. The GEMS signal curve continues smoothly upward.

Why would the GEMS signal maximum trail the fold-opening maximum? It is possible that as the folds begin the process of closing, they do not impede the airflow significantly. However, by closing slightly, the glottal area is somewhat reduced, increasing the Bernoulli force acting radially inward on the folds and tracheal wall. This force would tend to pull the tracheal wall inward even further. As the closing of the folds progresses, the resistance of the glottis would increase enough to impede the airflow and the subglottal pressure would begin to rise. The tracheal wall's motion inward would begin to slow under this higher pressure. Indeed, as the closing occurs, the position of the wall (radially inward) begins to slow and when the folds fully seal (position 11 on

Figure 3.10) the position changes abruptly. Thus, the GEMS-derived position matches the expected motion of the tracheal wall due to subglottal pressure changes.

In summary, for this portion of the signal we are able to detect fold closure but are not able to detect the large nodule on the right fold as it contacts the left fold. This would indicate we are seeing the movement of something other than the folds, which is nonetheless closely tied in with fold closure, such as sub- or supraglottal tissue vibrations.

Before, after, and at the minimum of the GEMS signal (folds closed)

After its rapid decrease (point 12 in Figure 3.10), the GEMS signal starts to bottom out and smoothly reaches its minimum value about 0.8 msec later. From point 11 in Figure 3.10 until about 0.2 msec from the minimum the folds are firmly closed. From Figures 3.9 and 3.10 it can be seen that there is a significant change in the GEMS signal when the folds are not in motion. This is another indication that the folds do not directly affect the GEMS signal.

Conclusions for abnormal physiology

1) There is no discernable change in the GEMS-derived position, its derivative, or its integral to indicate the nodule has broken or initiated contact with the left fold. This is strong evidence that the folds are not affecting the GEMS return directly, either through transmission or reflection.
2) The point of maximum cross-sectional glottal area occurs approximately 0.25 msec before the maximum of the GEMS signal. Thus the folds have reversed direction without any indication from the GEMS that they have done so. This indicates that the GEMS is not measuring fold motion. On the other hand, the GEMS signal is

consistent with that of the position of a vibrating tracheal wall driven by changing subglottal pressure.

3) The rapid drop of the GEMS signal associated with fold closure occurs after almost all of the fold area is already in contact. That such a small change in contact area would have such a large effect on GEMS return is unlikely if transmission or reflection of the GEMS from the folds had occurred.

Analysis of normal physiology

In this section, we will examine normal physiology folds (with the author as the subject) with a frame rate of 3000 fps. The exposure time was reduced to just 30 μsec resulting in clear, sharp pictures.

There were two modes of voicing in this experiment – falsetto, then a transition to modal. The falsetto fundamental was approximately 300 Hz and the modal about 150 Hz. I will examine the results only for the falsetto, which due to its higher frequencies is mostly unaffected by the filters of the GEMS. The results for modal speech were identical to those for falsetto.

During the falsetto register the folds are held under greater tension and do not make as much contact as in modal or fry. In fact, the duty cycle (% of time the folds are open) is about 85, so the folds are only closed for a short part of the total period. However "closed" is a bit of a misnomer, as in this part of the experiment the folds never quite achieve full closure, always exhibiting a posterior "chink" (see Figure 3.11). This could allow air to flow from the subglottal to the supraglottal region even when the folds are closed, reducing the subglottal pressure. Thus the subglottal pressure cannot rise as high due to the significant leakage during closure. This may be part of the reason the

amplitude of the GEMS signal is lower (it is about one-third) for falsetto than modal. Another factor is the higher frequency of vibration (about 300 Hertz), which does not allow the tracheal wall to change position significantly in the relatively short time available. A third factor arises from the physiology of producing falsetto speech. During falsetto, the cricoid and thyroid cartilages rotate with respect to one another, changing the orientation and location of the tissues in the glottal area. Thus the distance to the reflector (the tracheal wall) may change and a different sensitivity could cause the change in amplitude. Figure 3.12 summarizes the observations I have made for the falsetto section of the normal physiology.

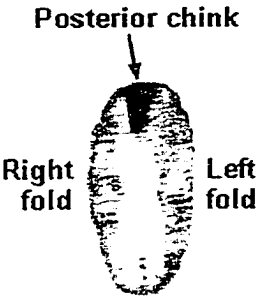


Figure 3.11. Example of "fully closed" folds in falsetto mode

Before minimum (at rapid drop) of GEMS signal

The first thing I noticed was that the rapid drop in the GEMS signal (just after frame 91 in Figure 3.12) does again not correspond with the first touch of the folds. The first touch of the middle of the vocal folds occurs about one half of a frame (about 1/6 of a millisecond) before the drop. There is again no GEMS signal associated with this contact. As with the abnormal physiology, the drop itself does not occur until the anterior portion of the folds begins to close and the glottis becomes quite small. This is understandable if we remember that the glottal resistance is approximately proportional to glottal width^{-4}, and so the glottal resistance (and resulting subglottal pressure) could rise sharply even though the folds are not yet fully closed.

At the minimum of the GEMS signal the anterior is fully closed, but as discussed earlier the posterior is always open slightly. The anterior portion of the folds begins to

open within about one half of a frame (~ 1/6 msec) after the minimum. The folds open from the anterior and posterior portions first and the middle portion of the folds is last to open. It is significant that the GEMS signal begins to rise before the folds begin to open. Again, the GEMS indicates a change in position when the folds are closed. Thus again the GEMS-derived position does not correspond to fold motion, but rather to subglottal pressure driven tracheal wall motion.

Soon after the linear part of the GEMS return begins (after frame 93 in Figure 3.12) the anterior portions of the folds opens. By the middle of the rise, the anterior and posterior sections are both open. However, the middle of the folds does not separate until after the elbow at the top of the rise. Thus the large increase in fold transmission resistance is not accompanied by a significant change in the GEMS return. Thus again, transmission through the folds is unlikely.

In the next three hundred microseconds (between frames 94 and 95), the glottal area expands rapidly and the folds are moving very quickly. But the GEMS-derived position, in contrast, is slowing. Therefore, as before, the fold motion is quite different from the GEMS signal.

The region of maximum glottal area for the falsetto again occurs about 250-300 microseconds before the maximum of the GEMS signal is encountered. At the maximum of the GEMS signal the folds have already begun to close, although the middle of the folds doesn't touch until about 2/3 of a frame before the rapid decrease in the GEMS-derived position begins. Again, as in the opening of the folds, there is no event signifying fold closure as would be expected if the folds are directly related to the GEMS signal.

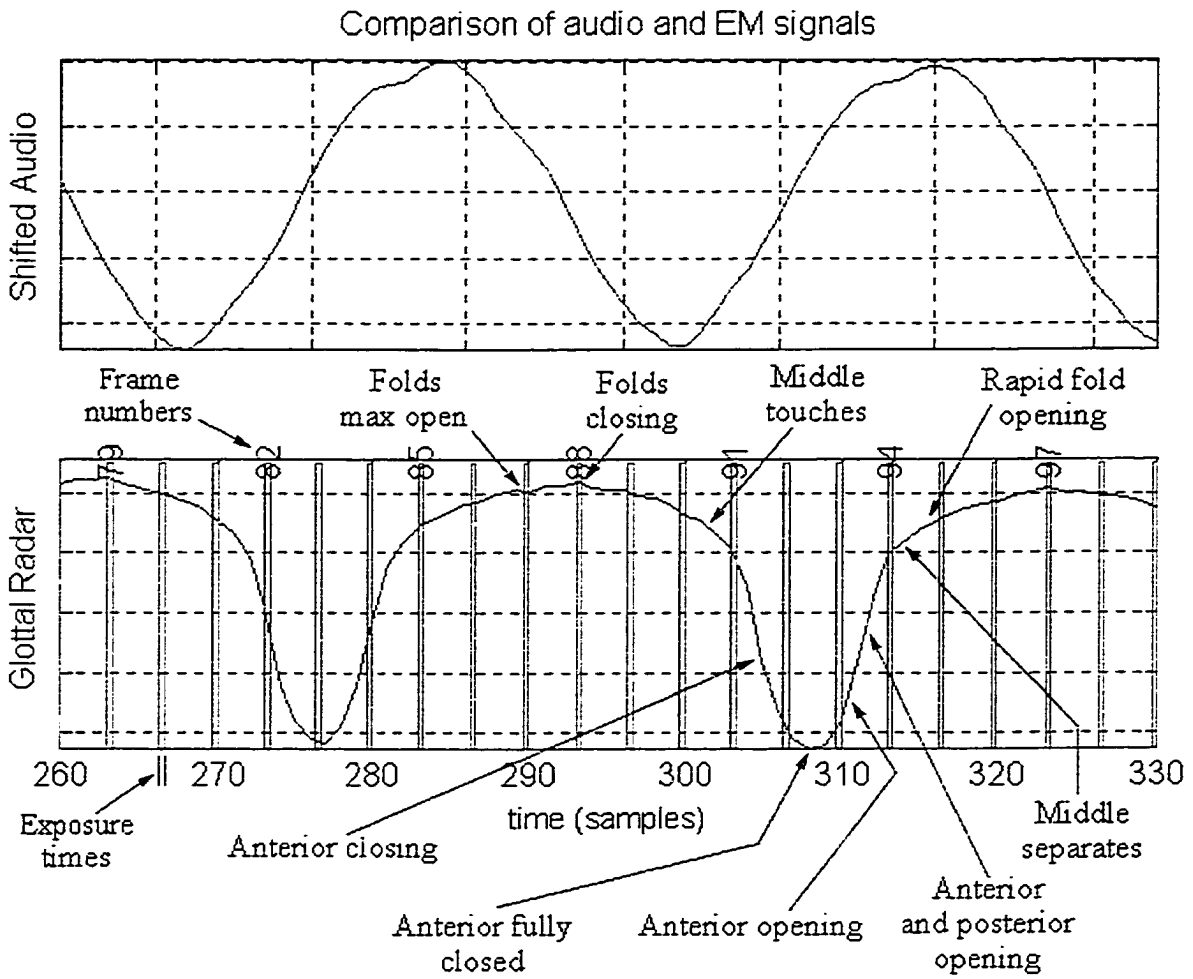


Figure 3.12. Summary of observations for the falsetto portion of the normal physiology. The time scale is in samples, at 10000 samples/second.

The results for normal physiology modal vibration are basically the same as in the previous two sections, for the sake of brevity the observations are omitted and only the conclusions discussed.

Conclusion for normal physiology

1) The first and last touch of the folds does not correspond to any significant GEMS signal. As with the abnormal physiology, the GEMS-derived position data does not correspond to fold motion, but instead matches well what would be expected for tracheal wall vibration.

2) The rapid drop in GEMS signal magnitude is associated with the closure of the anterior portion of the glottis and the corresponding increase in subglottal pressure.
3) After the minimum of GEMS signal is reached, it begins to increase before the folds have reopened. This is due to the slowing of the motion of the tracheal wall as the subglottal tract pressurizes. Once again, the GEMS signal does not correspond with fold motion.
4) The regions associated with rapid fold movements (either opening or closing) are associated with quite placid GEMS signal. Conversely, the regions associated with large changes in pressure and airflow are accompanied by small fold position changes, but large GEMS signal magnitude changes.
5) As before, the region of maximum glottal area and the point of maximum GEMS signal do not coincide. Thus the folds are closing as the GEMS magnitude continues to increase. Again, the GEMS signal has little to do with fold motion.

Answers to questions posed

1) Consistently, in all experiments, the GEMS signal was shown conclusively to deviate from observed fold motion. This occurred at the maximum of the GEMS signal (the folds were closing), at the rapid decrease (the folds were almost closed before the rapid drop-off in GEMS signal occurred), at the GEMS signal minimum (the GEMS would consistently begin to rise before the folds opened), and during the rise of the GEMS signal from its minimum (the anterior, posterior, and middle sections of the abnormal folds were in constant motion and contact during this time, but the GEMS signal was quite smooth). In addition, the GEMS-derived position signal matches quite well to the motion expected from the tracheal wall driven by

pressure changes. Thus the evidence presented here argues quite forcefully for a tracheal wall motion basis for the GEMS return, the trachea being driven by either the sub- or supraglottal pressure.

2) In the abnormal case, we were not able to directly detect the contact of the nodule and the opposite fold. As the nodule was about 1/3 the length of the fold, the fact that there was absolutely no GEMS signal change at contact is significant. It is further strong evidence against reflection from the folds or transmission through them.

Section 3.4. The University of Iowa experiments

On September 19th and 20th, 1996, Drs. Ingo Titze, Brad Story, Wayne Lea, and I conducted several experiments at the University of Iowa National Center for Voice and Speech's anechoic chamber. They involved the use of an electroglottographic (EGG) device, a condenser microphone, and the GEMS. Data were taken on all four participants, in various registers, pitch, and intensities, for the three vowels /a/, /i/, and /u/. This resulted in a great wealth of data, which we are still finding instructive.

The EGG is an alternating current (in the MHz range) device that uses a bistatic configuration to measure the conductivity of the tissue between the electrodes. It uses two electrodes, a transmitter and receiver, which are placed on opposite sides of the trachea near the folds. Wherever the folds are in contact, there is conduction of the current, if there is an air gap, there is no direct conduction. Thus, the strength of the signal gives a good representation of the amount of vocal fold contact area. There is no scattering from tissue interfaces because the electric probe current does not propagate as a wave in the tissue. The conducting electrons and ions are moving over very short mean free paths; thus the mode of propagation is conductive rather than transmissive. As the GEMS and the EGG operate in different parts of the EM spectrum, there is no interference in operation between the two.

The vertical and horizontal positioning of the electrodes are critical to EGG signal strength – they must be on either side of the folds so that they may use them to cross the trachea. EGG signals are sometimes difficult to obtain from women (due to their more recessed tracheas) and those with a large amount of neck tissue or fat (which allows more

of the RF signal to bypass the folds). For more information on EGG operation, see Titze (1984).

Procedure: For most experiments, the subjects phonated in a variety of registers, intensities, and pitch, with the EGG and GEMS data recorded simultaneously. However, one set of experiments was performed with only the GEMS. In this set the GEMS was moved up and down the trachea, as well as side to side, in 1 centimeter increments from the laryngeal prominence. The outer case of the GEMS was kept perpendicular to the skin at all times. A recognizable signal was obtained as far as 6 cm to the side of the prominence, and as far up and down as it was physically possible to go (2 cm and 6 cm, respectively). In fact, very small (but still recognizable and of the correct frequency) signals were obtained from various places on the head during speech, probably due to miniscule skull vibrations due to vocalizing. Data was recorded on a Sony PC-108M DAT recorder at the maximum sampling rate of 10 kHz.

Objectives:

This experiment will deal with the inverse filtered GEMS signal (the position of the reflecting tissue interface), as the intent was to compare the GEMS and the EGG signals to see if they are correlated. Since the EGG signal is directly related to vocal fold contact, if the signals had similar characteristics then vocal fold contact may be a factor in the GEMS signal. At this date, the high-speed video experiments had not been performed, so the strong evidence against a fold motion component of the GEMS signal had not yet been discovered.

3.4.1 Analysis: Comparison of GEMS and IEGG

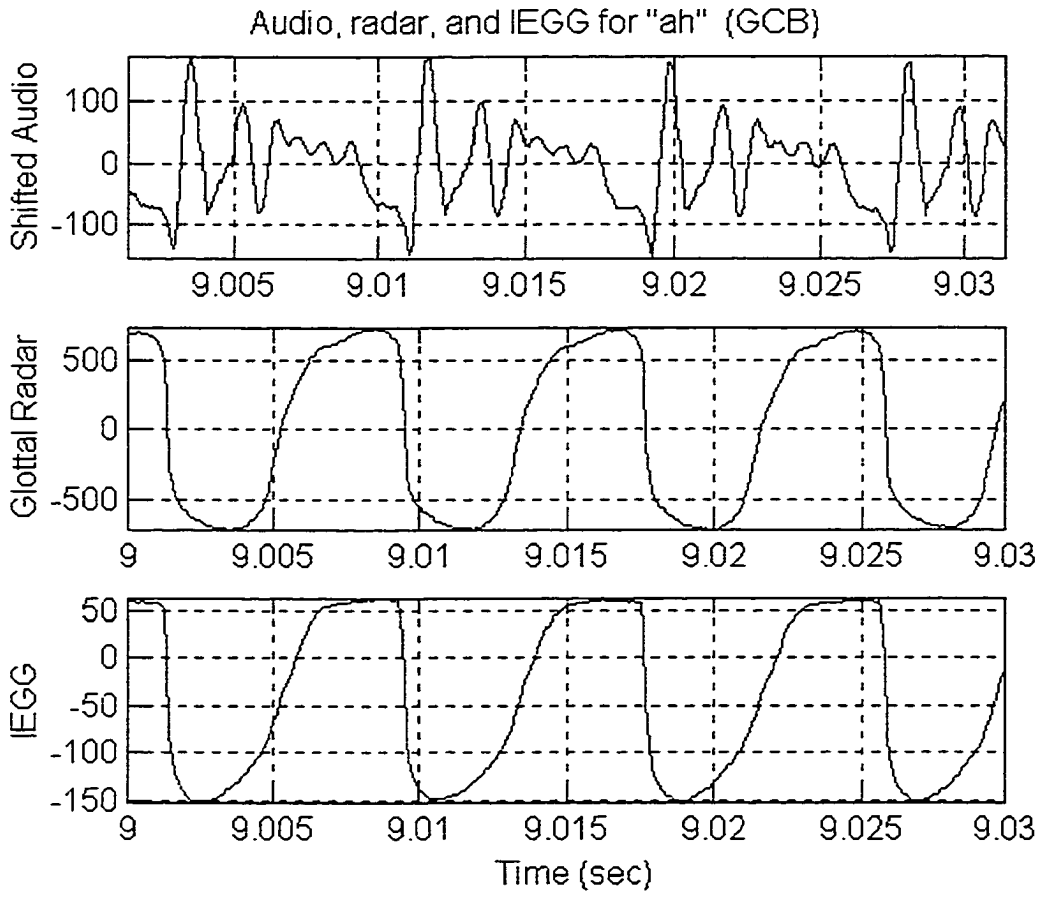


Figure 3.13. Audio, GEMS, and inverted EGG (IEGG) for /a/.

Upon first examination, the GEMS and EGG look quite similar for modal speech (Figure 3.13). Here, the shifted (by 1.3 milliseconds to account for the travel time from the glottis to the microphone) audio is plotted on top, the GEMS is in the middle, and the inverted EGG (IEGG) signal is shown on the bottom. The EGG is inverted to facilitate comparison between it and the GEMS. Inverted, a positive signal indicates very little conductivity (folds open), and a negative signal indicates large conductivity (folds closed). Coincidentally, the similarities between the inverted EGG (IEGG) and the inverse filtered GEMS are quite striking. Except for the concavity of the ascending line (between about 9.021 and 9.023 seconds) and some differences near the negative peak,

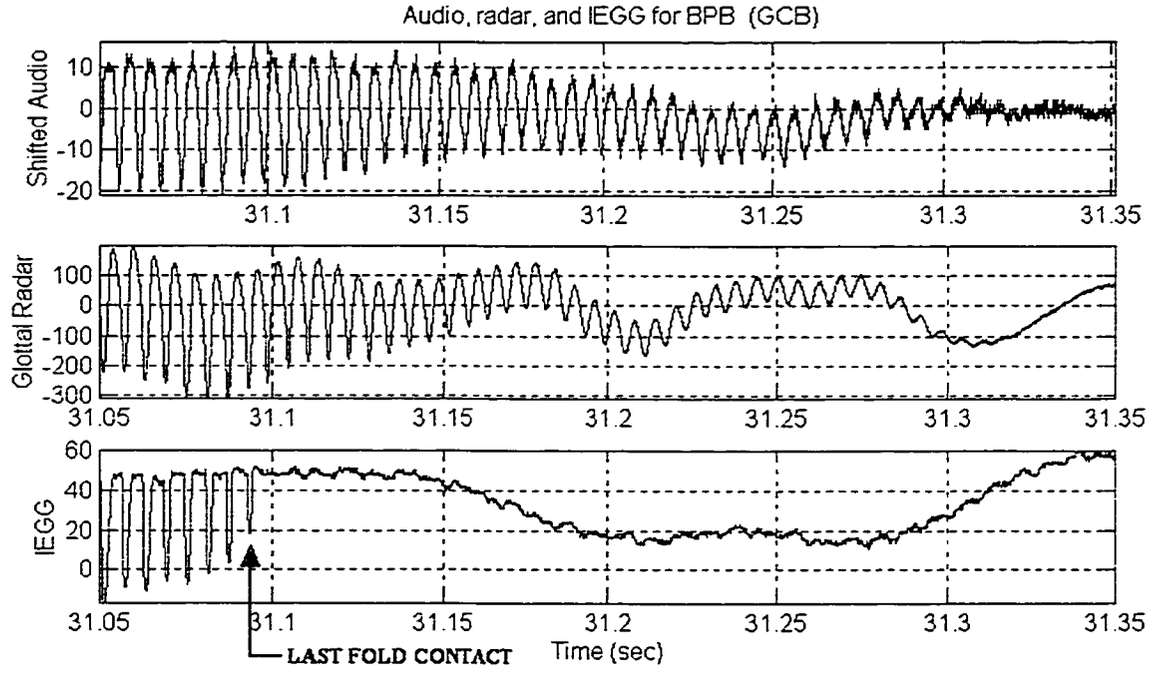


Figure 3.14. Plot of audio, GEMS, and IEGG for breathy cessation of speech. Note the total lack of EGG signal as contact is lost, and also the similarity of the audio and GEMS near the end of the speech.

the GEMS and EGG signals are nearly identical. This was one of the reasons that led to the transmission theory (hypothesis 2) discussed in the introduction to this chapter.

Unfortunately for the transmission theory, the IEGG and GEMS signal disagree in many important ways. One of the most graphic is the difference noted when beginning or ending speech in a breathy (i.e. low phonation pressure) voice (see Figure 3.14). At the beginning (end) of breathy voice, the folds are pushed together (pulled apart) gently, and fold vibration without contact can occur for many glottal cycles. The airflow is still being modulated, and sound is still being produced, but there is no vocal fold contact. When this occurs, the EGG does not register any signal at all, as its current only flows through tissue, not air. However, in Figure 3.14, the GEMS registers a large signal that does not change appreciably in magnitude or shape with the loss of contact. This is an

unmistakable sign that transmission of the GEMS wave through the folds is not occurring.

This fact that the composition and size of the signal just before and after contact remains relatively constant is another clue to the origin of the GEMS signal. There are many, many instances of this occurring in the data. If we were actually detecting the folds directly in some manner (either through transmission or reflection), then contact would certainly be noticeable. If we are detecting tracheal wall motion, the signal would not change substantially as the change in glottal area (and therefore resistance) between co-located contact to non-contact cycles would not be substantial (see Section 2.3.2 for a glottal resistance discussion).

Another interesting situation that was observed consistently was that the GEMS signal tended to mirror the audio signal quite well in phonation situations without contact. For example, see the GEMS and audio signal in Figure 3.14 from about 31.2 to about 31.3 seconds. Without fold contact, reflections from the folds would be at a minimum (the glottal region is no longer acts as a close-ended pipe, leaving a relatively clear path through the folds and into the lungs). Therefore any reflections of the audio from the mouth/air interface that propagate back to the folds would tend to pass through them and into the lungs, which are very lossy and do not support reflections. This means that the audio escaping the mouth should be a good representation of the subglottal pressure pulse, as there are few reflections to enrich the signal. If the GEMS does represent the position of the subglottal tracheal wall, a close resemblance to the recorded audio during this time would be expected. Indeed, in Chapter 4 we will show that the audio and the GEMS-derived pressure are similar in this instance, especially in phase. Conversely, if

the GEMS signal depends on fold motion or transmission, it would be quite a coincidence to match the audio in both shape and relative magnitude so well.

Finally, there is the difference in signal strength when the IEGG and GEMS are moved vertically and horizontally from the laryngeal prominence. The EGG signal disappears completely on the order of a centimeter up or down, and within a few millimeters left and right. The GEMS, on the other hand, detects signals (not that much different than the signals it detects at the glottis) up and down the entire trachea, and as far left and right as 6 cm. This data will be examined in the next section.

3.4.2. GEMS position experiment analysis:

As noted above, we conducted several experiments in which the GEMS was moved horizontally and vertically from the laryngeal prominence in 1 cm steps. This allowed us to observe the changes in signal as this occurred, in hopes of gaining insight into how the GEMS operated.

Horizontal data set:

The first set of data to be analyzed from the position experiments were the data in which the GEMS was moved to the left (Figure 3.15) and right (Figure 3.16) of the laryngeal prominence in 1 cm increments. From the data in these two experiments (the other subjects' results were similar), it can be seen that the signal decreases as the GEMS was moved away from the central point. There is also a 180 degree phase change at 4 cm moving to the left and again at 3 and 5 cm moving to the right. This is expected due to the increase in effective path distance to the trachea. The increase in effective path length causes the GEMS to operate on a different part of the sensitivity envelope discussed in

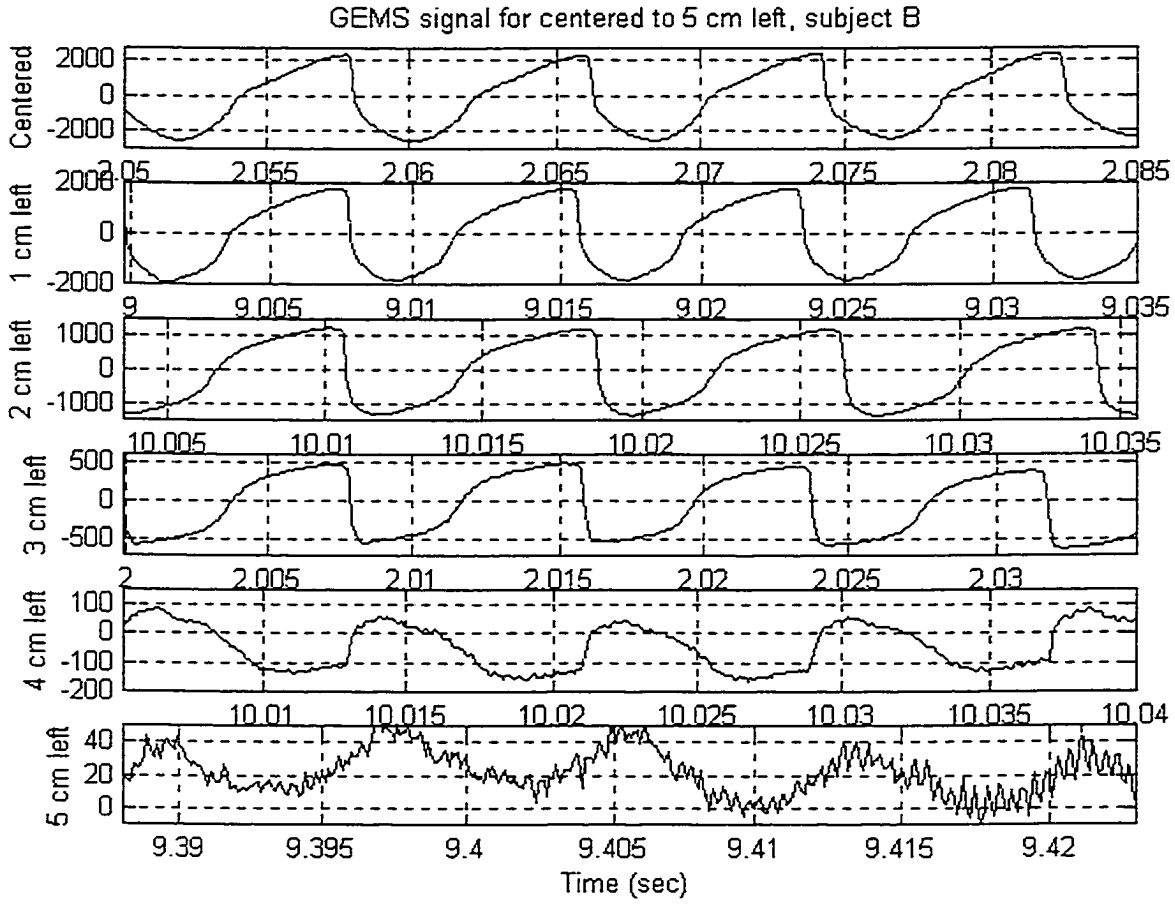


Figure 3.15. Data from position experiments when GEMS is moved from the center of the trachea (the laryngeal prominence) to 5 cm to the left of the prominence. Note the phase change at 4 cm.

Section 2.1.3. The change in phase denotes that we have moved across a null of the envelope.

Notice, too, that the sharp feature of the GEMS signal is consistently present, up to 4 cm to the left and even barely at 5 cm to the right, although it is quite faint there. These observations are all completely supportive of reflections from the tracheal wall. As we move the sensor more than a few centimeters to the side of the prominence, we are most likely scattering off the side wall of the trachea, not the anterior or posterior portion. Still, with our knowledge of the sensitivity envelope and reflectivity of the almost-round trachea, this data is completely consistent with tracheal reflection.

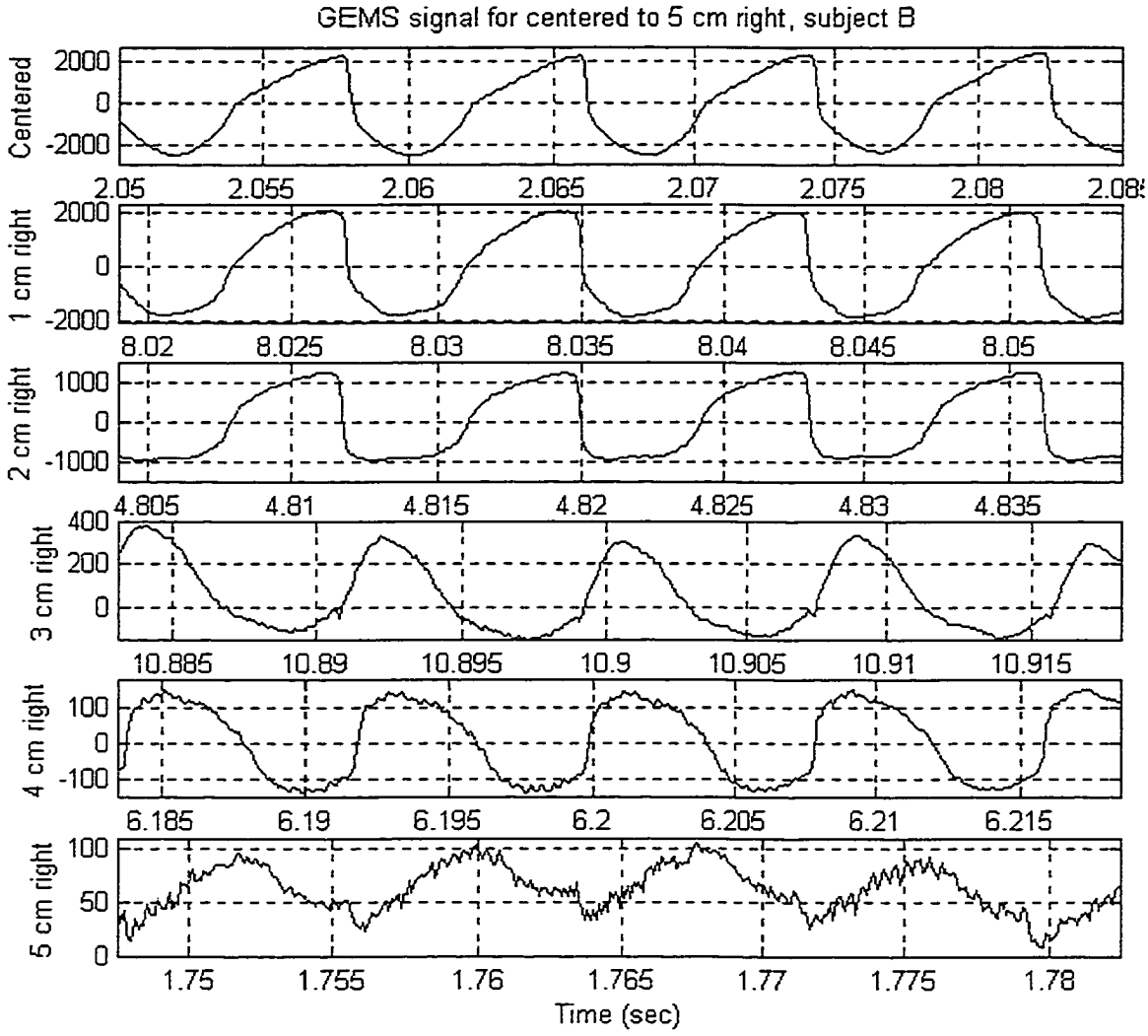


Figure 3.16. Data from position experiments when GEMS is moved from the center of the trachea (the laryngeal prominence) to 5 cm to the right of the prominence. Note the phase change at 3 cm and again at 5 cm.

This decrease in amplitude of the GEMS signal as a function of azimuthal angle around the neck has been suggested as evidence for transmission through the folds and against reflection from the folds (without considering the trachea as a possible reflector). The reasoning is this: If EM wave transmission is occurring (from the GEMS, through one fold, across to the other fold, then back around to the GEMS) when the GEMS is located at the front of the neck, then as the GEMS is rotated around the neck, to the side, the amplitude of the GEMS signal should drop, as there is no way for the transmitted EM

signal to return to the GEMS receiver. This is precisely what occurs with the EGG. Moving the electrodes only a centimeter or two to the side reduces the signal completely to noise. On the other hand, if reflection from the medial fold surface is occurring, than the GEMS signal amplitude should grow stronger as we move the GEMS to the side of the neck, as more of the medial surface (assumed to be the reflector) comes into view. We have already shown that the GEMS does not reflect well from the folds (either from the front or the side) due to a combination of geometrical factors and diffraction, but it is important to examine the anatomy of the neck to show why the observed data is consistent with the tracheal wall reflection model rather than the transmission model.

A review of the anatomy of the glottal area (see Figure 3.17, a slice of the visible human, the transverse view at the level of the vocal folds, is presented) will show that the neck is not round, but an oval with its long axis oriented from left to right. Furthermore, the trachea is not at the center of the neck, but rather the spinal column occupies that

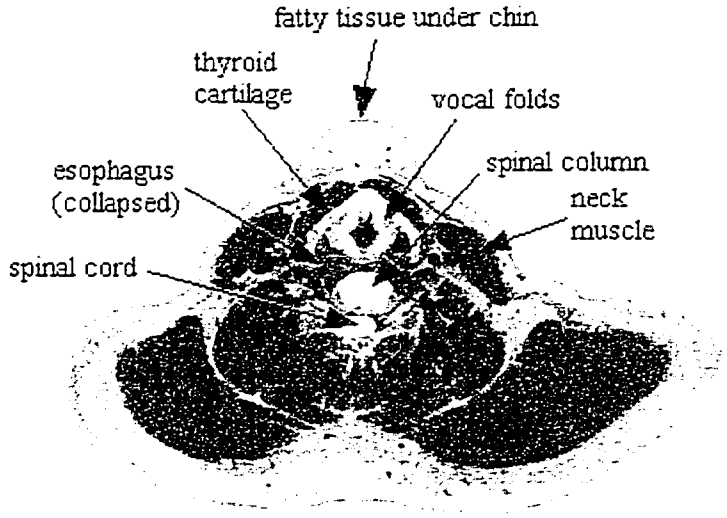


Figure 3.17. Slice 1250 of the visible human (available at http://www.npac.syr.edu/projects/vishuman/VisibleHuman.html)

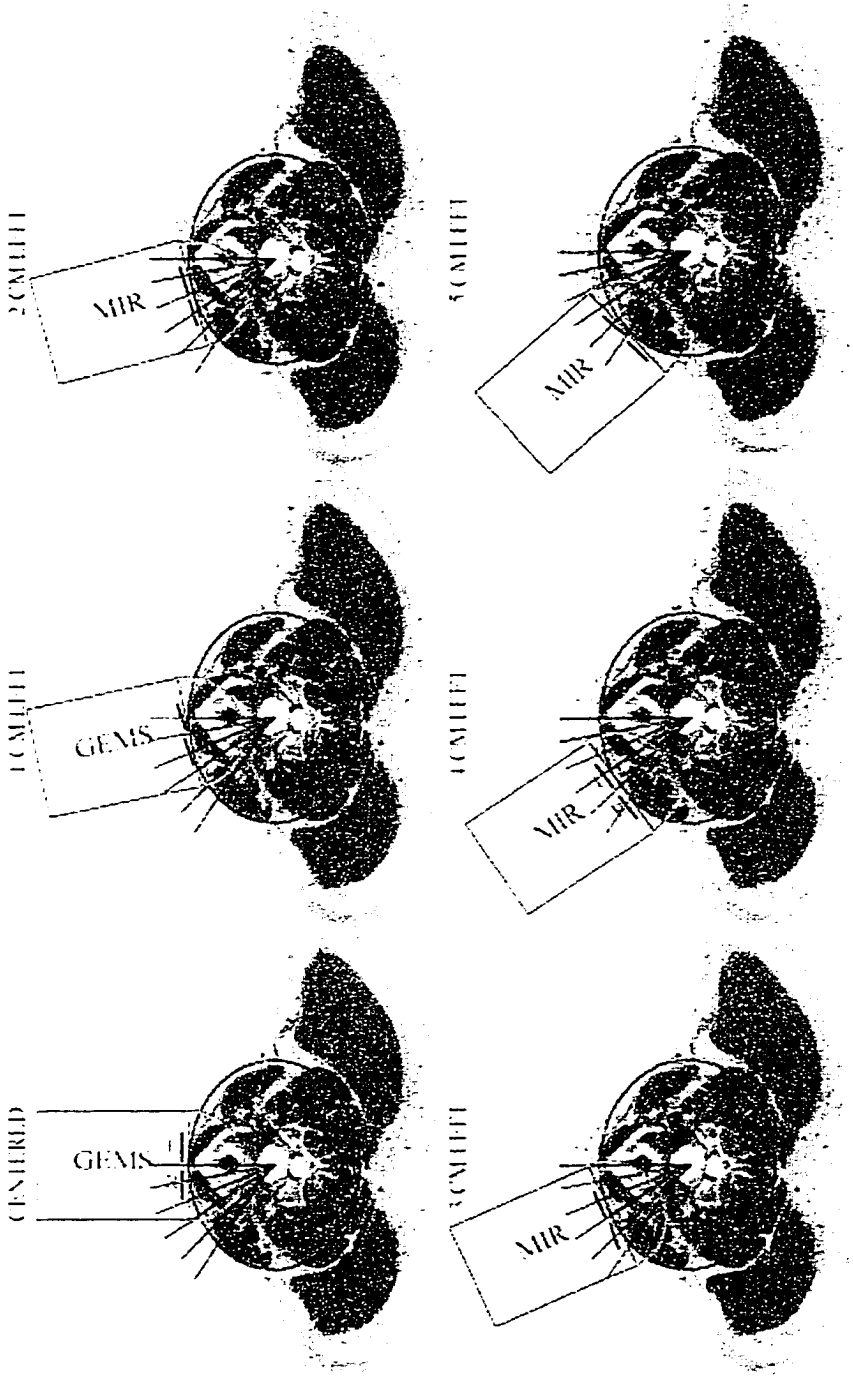


Figure 3.18. The visible human slice and the GEMS, moved 1 cm at a time to the left.

position. In Figure 3.18, we see several views of the visible human with an appropriately scaled GEMS box as it traverses the neck at 1 centimeter intervals. It is kept perpendicular to the skin as it moves to the left, as was the procedure in the experiments. As the azimuthal angle increases, the GEMS is no longer oriented toward the folds. The oval shape accentuates this occurrence, as the GEMS is held perpendicular to the neck as it is moved outward from the center. Even assuming for the moment that transmission from one antenna, through the folds, and back to the other antenna occurs when the GEMS is in its normal position, as the GEMS is moved to the side transmission looks quite improbable after only 2 or 3 centimeters. At 2 cm the antennae of the GEMS are both on one side of the neck, aimed at the same fold. Yet the GEMS signal is still quite strong at that distance.

The transmission theory also neglects the equivalent free-air path length from the antennae to the folds/trachea along with the sensitivity envelope. Since the distance to the reflector increases as we increase the azimuthal angle between the GEMS and the centerline of the neck, we move along the sensitivity envelope. Recall from Section 2.1.3 that the distance between nulls is about 35 mm in air. In the neck tissue, composed of high dielectric muscle with an index of refraction of about 7, that distance is reduced to about 5 mm. We do observe changes in the magnitude and sign of the GEMS as we increase the azimuthal angle, so it is clear we are moving along the sensitivity envelope and the GEMS magnitude will vary accordingly.

Another factor that causes the signal strength to drop at high azimuthal angles is the losses in the EM wave due to the conductivity of the medium. It is clear from Figures 3.17 and 3.18 that there are many muscle groups surrounding the trachea, and as the

azimuthal angle increase, the amount of tissue through which the wave must propagate grows substantially. In Figure 3.18, an orange semicircle of constant radius has been drawn around the GEMS. Note how the distance increases substantially after moving 3 cm to the side. At the center position, the wave has to travel about 15 mm from the transmit antenna, to the trachea/folds, and then to the receive antenna. Displaced 5 cm to the right, the round-trip distance to the nearest part of the trachea has increased to ~70 mm, and the distance to the nearest medial surface of the folds is has risen to ~114 mm. Recall from Section 3.2 that the skin depth (the distance at which the EM energy has been reduced by 1/e) in muscle tissue is about 18 mm. Thus, at the 5 cm mark the GEMS signal energy will be reduced to about $1/e^{3.9} = 2\%$ for the trachea, and about $1/e^{6.3} = 0.2\%$ for the folds.

Thus the drop in amplitude of the GEMS signal is actually expected for the reflection model, as well as the transmission model. The fact that there is any signal at all after moving 2 cm to the side supports the reflection model. From the accompanying data in Figures 3.14 and 3.15, we see that at 4 cm to the left the amplitude has decreased to about 5% of that observed at the center, leading to a reduction in energy of about 0.25%. This is certainly reasonable given the energy losses due to conduction and the movement along the sensitivity envelope. How much of the change in magnitude is due to the increasing distance and how much is due to the change in sensitivity is not known. Remember, the magnitude and polarity of the GEMS is not constant for a constant amplitude of motion as the distance between the GEMS and the reflecting interface is changed. Therefore, unless we are certain of the equivalent free-air distance of the GEMS from the reflecting surface, and thus know what the sensitivity is at that point, we

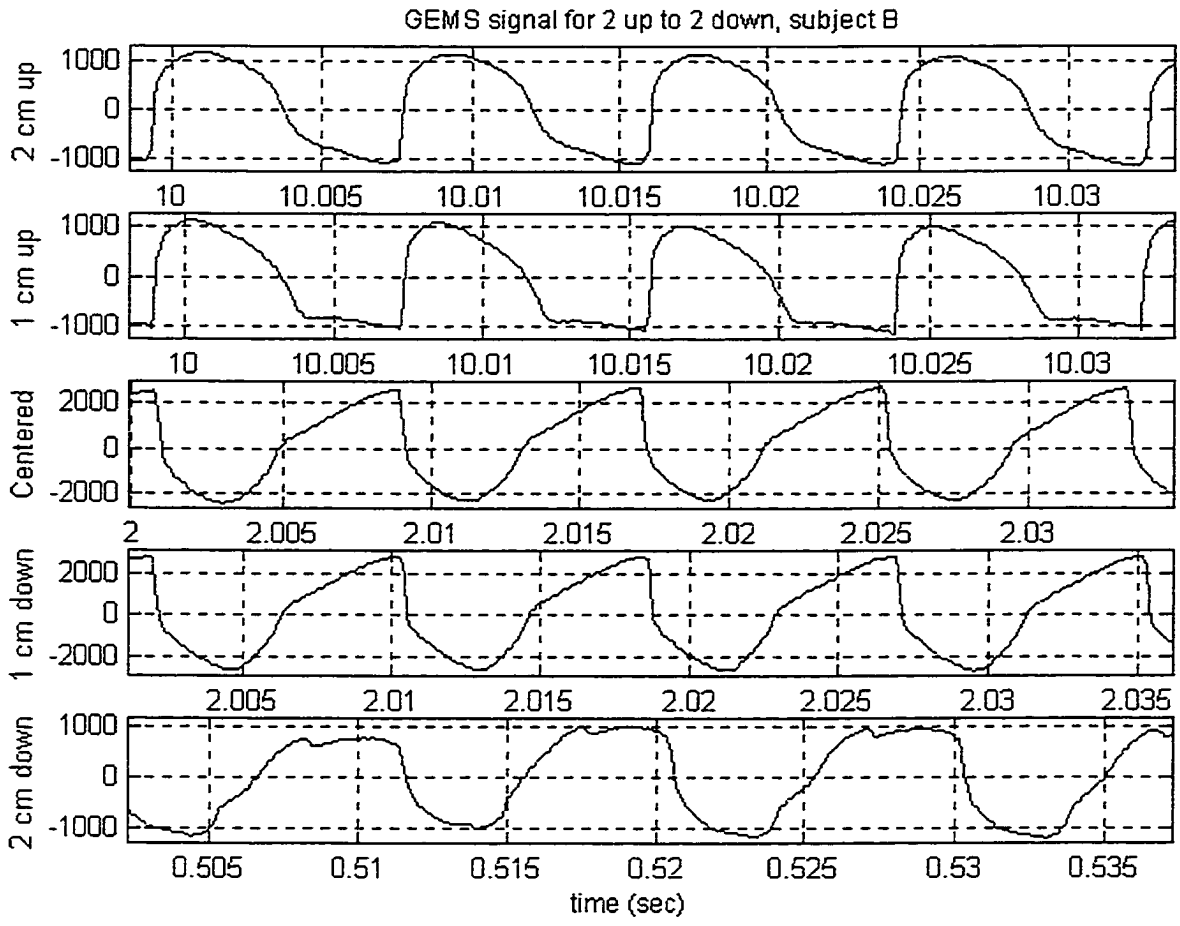


Figure 3.19. Data from position experiments when the GEMS is moved from 2 cm above the laryngeal prominence to 2 cm below it. Note the phase change from the positions above to the center.

cannot determine to what extent the energy loss is due to conduction or a reduction in sensitivity.

Vertical data set:

The data derived from moving the GEMS up and down the trachea was quite illuminating. The representative data from 2 cm above the prominence to 2 cm below it from subject B (again, all subjects' results were similar) is shown in Figure 3.19.

The most interesting thing in this figure is the phase change between the data above the prominence and the data at or below the prominence. This phase change was observed in every subject's data, and is direct evidence for tracheal wall vibration. The

reason for this is that the folds are attached to the thyroid cartilage just below the laryngeal prominence – which is defined as the vertical position of the glottis. Above the prominence is supraglottal tissue, below it is subglottal. If we are measuring either anterior or posterior tracheal wall motion in front of the GEMS antennae, then, when placed against supraglottal tissue (above the prominence) supraglottal pressure-induced vibration should be detected, and when placed against subglottal tissue (below the prominence) subglottal pressure-induced vibration should be detected. Subglottal and supraglottal pressures are roughly 180 degrees out of phase, as we shall see in the next chapter. If, on the other hand, the GEMS is measuring fold motion or transmission, there is absolutely no reason why a change in the position of the GEMS by 1 centimeter would invert the signal. Therefore the fact that the GEMS signal inverts when going from supra- to subglottal regions is direct evidence that the GEMS is measuring tracheal wall motion, not fold motion or transmission. It is also evidence that the normal placement of the GEMS at the laryngeal prominence results in the detection of subglottal rather than supraglottal tracheal wall motion.

Conclusions

In this experiment we learned that while the GEMS signal can look very similar to an EGG signal during normal (modal) speech, its behavior is quite different during breathy speech when there is no fold contact. There are also many other differences that were evident during almost any other vocal register or phonation level. We also reviewed experimental evidence indicating that the GEMS senses motion of the tracheal wall, rather than motion of the folds or transmission through the folds. The evidence for this includes the lack of any signal related to fold contact, lack of significant change to

the GEMS immediately before or after fold contact, the similarity of the audio and GEMS signals during periods of no fold contact, and change of phase of the GEMS when the fold region is transited vertically. Therefore, we may conclude that the GEMS is normally measuring subglottal tracheal wall motion.

Section 3.5. Anterior vs. posterior tracheal wall

In this section we will attempt to determine whether the position signal from the GEMS is originating from the front or the back of the tracheal wall. We will use an experiment in which the GEMS was slowly moved away from the neck, and we used the sensitivity null positions as reference markers. As the GEMS was moved away from the skin surface, the magnitude of the GEMS signal fell until a null in the sensitivity envelope was crossed and the polarity was observed to change. Continuing out away from the trachea, the GEMS magnitude rose slightly, then descended again until a second null was crossed. The distance (from the skin) for the null crossings was 20 and 65 millimeters. The magnitude of the GEMS signal was very near zero at the nulls, indicating that a single reflecting interface is responsible for most of the signal.

To ensure that the GEMS wave was penetrating the neck when the GEMS was pulled away and was not measuring surface skin vibrations, a thin copper foil barrier was glued to the exterior of the neck. It was quite light, so the skin would still be free to vibrate. As copper is a near-perfect reflector, so if any vibration of the exterior surface of the neck was causing the GEMS signal, the return should have been greatly amplified.

However, the amplitude of the GEMS was observed to diminish greatly, signifying that the great majority of the signal was being generated within the body. At approximately 5 millimeters from the neck, with the foil in place, the signal decreased in amplitude by a factor of 8.

The distances from the neck that the null points were observed should be able to tell us if the GEMS signal originates from the anterior or posterior wall of the trachea. We have seen in the previous section that the posterior wall (except at the cricoid cartilage)

looks more likely to be the origin of the signal due to its flexibility and geometry. We use the sensitivity curve calculated from the shaker experiments (Section 2.1.3), which gave us the locations (distance in air from the GEMS) of the null points (places where the GEMS signal changes sign). Since we have many tissue layers involved in this experiment (the skin, fat, muscle of the neck, cartilage, and smooth muscle of the trachea) we have to calculate the effective path length (EPL) through the tissue. This is the path length the GEMS experiences while traversing the tissue. It is different from the path length in air since tissue has a non-unity index of refraction.

To determine the thickness of each layer of tissue between the air outside my neck and the air inside the trachea, I subjected myself to a CT (computed tomography) scan, which uses X-rays to image 1 mm thick transverse slices of the anatomy of interest. The scan was performed at the UC Davis Medical Center, using a GE Advantage CT scanner. One of the slices from the scan (the one located at the most inferior part of the folds) is shown at the end of this section in Figure 3.20, and a close-up of the region of interest is shown in Figure 3.21. Note the 1-cm tick marks to the side of the scan for scale.

From this scan, I have calculated the numbers in table 3.3. The uncertainties in the indices of refraction are due to individual variation as well as disagreements between researchers (Duck 1990, Lin 1986, Haddad *et al.* 1997) and tissue types and preparation. The thicknesses were measured on three adjacent slices (1 millimeter apart) and then averaged.

The effective path length (EPL) in the tissue to the front wall, then, will be at a minimum $1.7(5.5)+2.4(3.2)+2.4(7.0)+1.7(4.2)+1.7(7.0)=52.9\,\text{mm}$ and at a maximum

Material	Thickness (mm)	Index of refraction
Skin	1.7	5.5-6.6
Fat	2.4	3.2-4.0
Neck muscle	2.4	7.0-7.5
Cartilage	1.7	4.2-4.7
Trachea muscle	1.7	7.0-7.5
Tracheal width	24.8	1

Table 3.3. The measured thickness of the tissue layers that are between the anterior tracheal wall and the outside of the neck and their estimated indices of refraction.

$1.7(6.6)+2.4(4.0)+2.4(7.5)+1.7(4.7)+1.7(7.5)=59.6\,\text{mm}$. The EPL to the rear of the trachea would be between $52.9+24.8=77.7\,\text{mm}$ and $59.6+24.8=84.4\,\text{mm}$.

Referring to Figure 2.14 in Section 2.1.3, we see that the distance to the front wall (53-60 mm) is in an area of negative sensitivity and the distance to the back wall (78-84 mm) is in an area of positive sensitivity. The front wall distance is also quite close to the null point between 65 and 70 mm, while the back wall is near the maximum of 85 mm. As a result, the magnitude of the sensitivity is about twice as high for the rear wall than the front wall.

To get the total EPL of the two nulls observed, we take the distance from the GEMS to the skin of the neck (the 20 and 65 mm discussed above) and add it to the distances calculated to the front and rear of the trachea. The total distances are shown in blue in figure 2.14. Adding 20 mm to the range of 53-60 mm, we see that the EPL to the front wall will now be between 73 and 80 mm. This places the front wall squarely in an area of moderate magnitude. Adding 20 mm to the range of 78-84 mm yields a total EPL

to the rear wall between 98 and 104 mm, which effectively straddles the measured null between 100 and 105 mm.

Similarly, adding 65 mm to the original 53-60 mm gives us between 118 and 125 mm for the front wall, and adding 65 mm to 78-84 mm results in an EPL of between 143 and 149 for the rear wall. Examining the sensitivity envelope in this region, there is a null between 140 and 145 cm, and a maximum at 120 mm. Thus the rear wall EPL is very near two nulls on the sensitivity envelope, and the front wall EPL is near two maxima. This is conclusive proof that the GEMS signal is caused by motion of the posterior wall of the trachea.

By examining the sign of the sensitivity envelope, we may further corroborate the evidence above. As the folds close, the subglottal pressure rises and the trachea expands. Therefore, the front wall moves toward the GEMS and the rear wall moves away from it. At this time the GEMS registers a negative-going position vs. time signal. Does this make sense given the location of the EPLs calculated above?

With the GEMS placed directly on the neck, the EPL to the front wall is about 56 mm. This places it in an area of negative sensitivity. In this area, for reflectors with a higher index of refraction than the medium ($n_2 > n_1$), a negative signal means a motion toward the GEMS. However, it must be remembered that for reflections where $n_2 > n_1$, there is a phase change of 180 degrees upon reflection. For materials where $n_2 < n_1$, there is no phase change upon reflection. Therefore, for scattering from the anterior wall where the tracheal muscle has a higher index of refraction than the air inside the trachea, the sensitivity curve must be inverted. This means a negative signal actually denotes motion away from the GEMS, opposite what is expected.

For the normal GEMS position (held against the neck), the EPL to the rear wall is about 81 mm. This places the rear wall near a maximum of positive sensitivity. Since $n_2 > n_1$, the sensitivity curve does not need to be inverted, so a negative signal denotes motion away from the GEMS. This is consistent with expectations and corroborates the sensitivity null evidence described above.

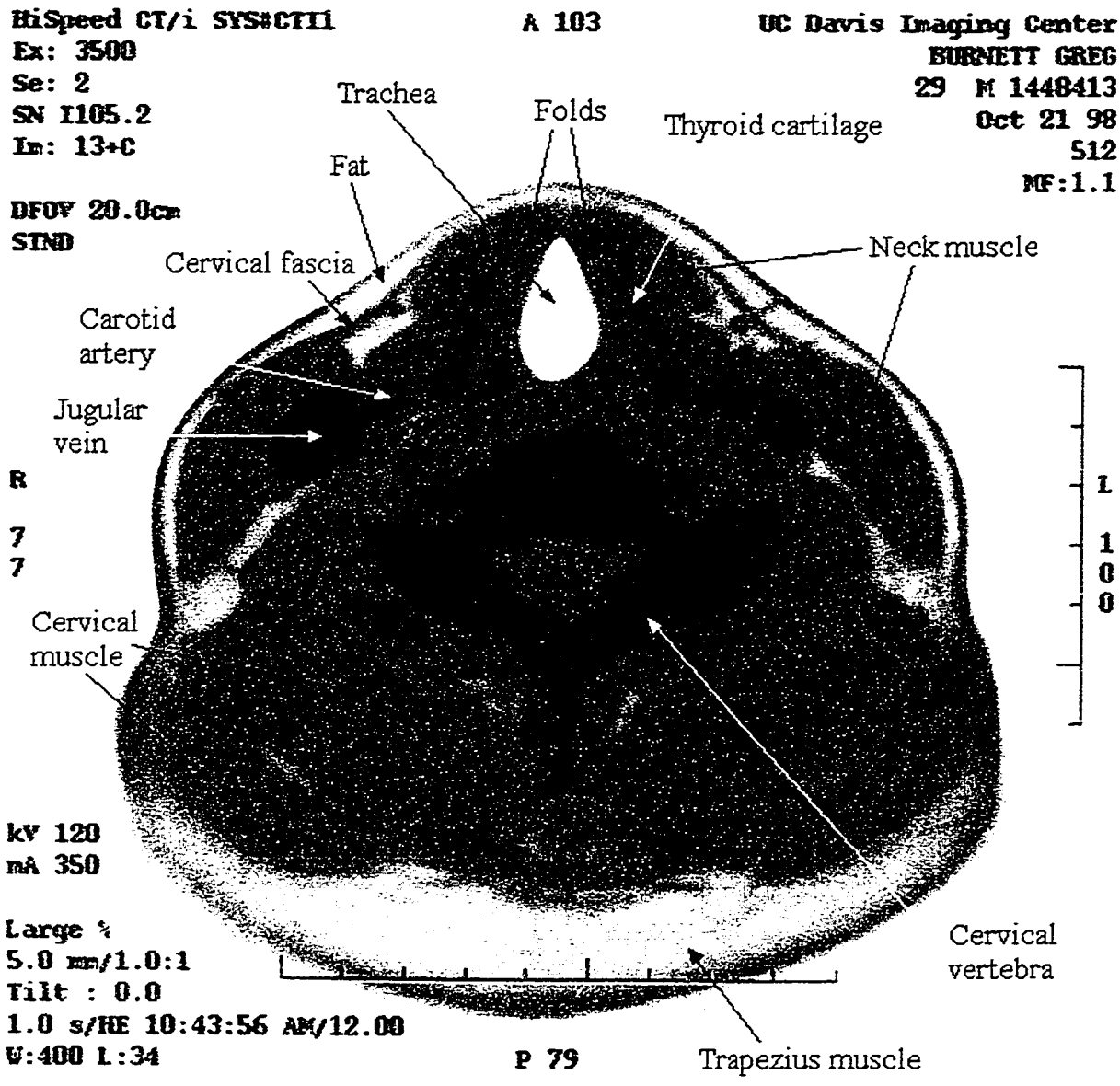


Figure 3.20. Slice from a series of CT scans performed on the author at the UC Davis Medical Center on October 21, 1998. Note the scale in centimeters to the right and below.

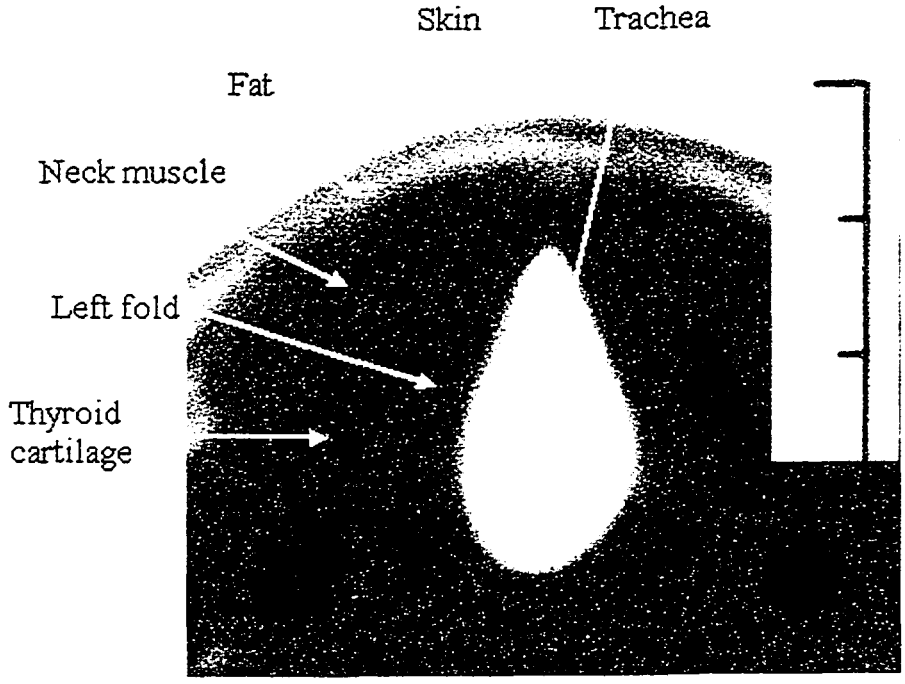


Figure 3.21. Expanded view of the region of interest form Figure 3.20.

Section 3.6. Conclusions on the physiological basis of the GEMS return

I have demonstrated several things in this chapter regarding the basis of the GEMS signal:

1) The signals observed come from the tracheal wall, not the folds (either by reflection or transmission). Sophisticated electromagnetic simulations demonstrated that the reflectivity of the tracheal walls would be at least 19 times as great as any from the folds. High-speed video experiments demonstrated that fold motion (both normal and abnormal) did not correspond to the GEMS velocity or position signals. The initiation or cessation of contact between the folds (which would surely cause a major disruption in the return if transmission or reflection from the medial line of the folds was occurring) went unreported by the GEMS.
2) Furthermore, in normal use the GEMS signals originate from the subglottal region. The evidence is this:
    i) The GEMS signal is still quite strong as the GEMS is moved down the trachea from its normal position. It even retains its sharp feature, and does not change in polarity.
    ii) The GEMS signal does change in polarity when moved above the normal subglottal position into the supraglottal range.
3) Finally, the posterior wall is responsible for the signal, and not the anterior wall. This was proven using two different procedures, using the calculated effective path length (EPL) through the tissues of the neck to the anterior and posterior walls of the trachea, the distance in air from the GEMS to the skin of the neck

where two nulls were observed, and the measured GEMS sensitivity curve from Section 2.1.3. It was shown that the polarity of the signal is incorrect for the front wall EPL, and correct for the rear wall EPL. In addition, the distance from the skin to the GEMS was measured when two sensitivity null points were observed. These distances were added to the calculated EPLs from the skin to the front and the rear wall of the trachea. This resulted in total EPLs for the front wall that were in areas of high sensitivity and total EPLs for the rear wall that were located very close to nulls. Therefore the nulls observed when the GEMS was moved away from the neck were caused by reflection off the posterior tracheal wall.

Now that we have determined that the posterior wall of the trachea is the physiological basis of the GEMS return, we will construct a model of the tracheal wall in order to get back to the driving force, the subglottal pressure. In turn, it will be used to define an excitation function for the vocal tract.

Chapter 4. Calculating the voiced excitation function

In this chapter, we will use the knowledge that posterior tracheal wall vibration is the basis for the GEMS return in an attempt to transform the GEMS return into subglottal pressure. In figure 4.1, a side view of the trachea is presented along with the airflow U(t), supraglottal pressure $P_{sup}(t)$, subglottal pressure $P_{sub}(t)$, and posterior tracheal wall position W(t). In order to derive the subglottal pressure $P_{sub}(t)$ from W(t), we must do the following:

1) Filter the W(t) using a digital filter that closely approximates a differentiator to get the wall velocity V(t).
2) Model the properties of a yielding wall with mass using a lumped-element circuit model.
3) Approximate the circuit wall model of the wall with a digital model.
4) Filter V(t) with the digital wall model to recover $P_{sub}(t)$.

We will begin by examining a lumped-element circuit model that models the effects of a yielding wall with mass.

Section 4.1. Converting velocity to pressure using a circuit model

To transform tracheal wall vibration into the driving force, subglottal pressure, we will need to construct a model of the tracheal wall. To fully understand the model of the tracheal wall, we will need to introduce the equations governing flow and pressure in a one-dimensional pipe of varying cross-section. A 1-D (or planar) approximation assumes that the acoustic pressure is the same across the entire cross-section at any point and therefore only the magnitude of the cross-sectional area (and not the shape of the area) is

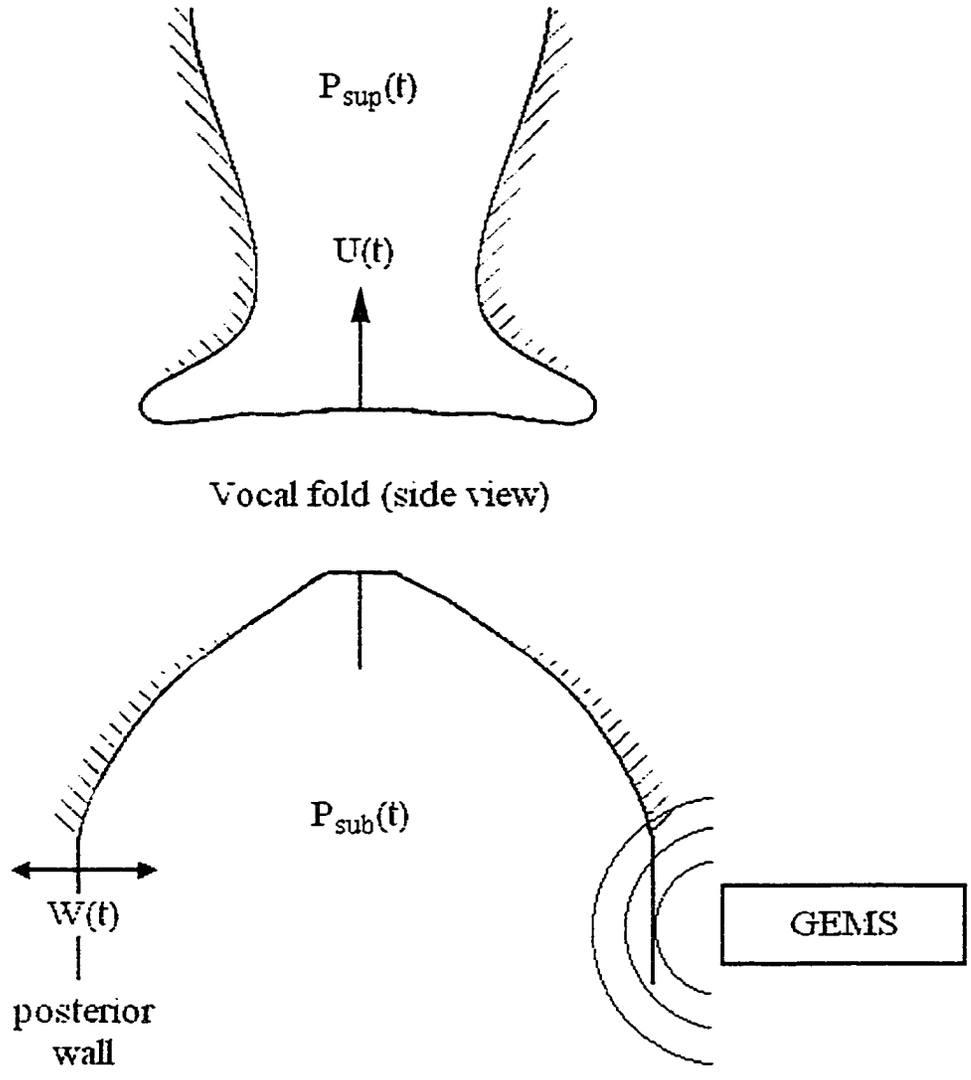


Figure 4.1. Side view of the trachea.

the important quantity. This approximation is only valid for frequencies below the cutoff frequency f_c:

$$f_c = \frac{0.5861 \cdot c}{2r}$$

where c is the speed of sound and r is the radius of the cross-sectional area under consideration (Eriksson, 1980). This equation assumes a circular cross-section, a good approximation for most of the vocal tract, especially the subglottal region. For a large

tract with a radius of up to 2 cm, the cutoff frequency is about 5100 Hz. The great majority of our energy resides below 3000 Hz, so the 1-D planar model may be considered sufficient under most circumstances.

Another approximation that is made in assuming 1-D flow is that the effects of the 90° bend at the base of the tongue on the flow and pressure are negligible. Story (1995) estimates that the resonances of the vocal tract are only affected a few percent by the presence of a bend in the tract, and the effects can safely be neglected. We will concur, as the effect is small and we are primarily interested in the subglottal region.

Introduction to the 1-D wave equation governing flow in a pipe

Assuming constant area in a small section, the relationship between 1-D flow and pressure are given by

$$\frac{\partial P}{\partial x} = -\frac{\rho}{A}\frac{\partial U}{\partial t} \qquad 4.1$$

$$\frac{\partial U}{\partial x} = -\frac{A}{\rho c^2}\frac{\partial P}{\partial t} \qquad 4.2$$

where A is the cross-sectional area, ρ is the density of the air, c is the speed of sound in the air, and P = P(x,t), the pressure; U = U(x,t), the volume velocity ($\bar{U} = A \cdot \bar{v}$) (Morse and Ingard, 1968). The one-dimensional wave equation can be obtained by differentiating 4.1 with respect to t and 4.2 with respect to x and then eliminating the common term to yield

$$\frac{\partial^2 P}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 P}{\partial t^2}. \qquad 4.3$$

The solution to this wave equation is the superposition of forward and reverse traveling waves $P^{\pm}(x,t) = C_{\pm}e^{j(\omega t \mp kx)}$, where $k = \omega/c$.

It can be shown (Kinsler et. al, 1982) that the above can be manipulated into

$$U^{\pm}(x,t) = \frac{A \cdot P^{\pm}(x,t)}{\rho c} \qquad 4.4$$

assuming ρ is constant and the amplitude of vibration is not too large (a good approximation for most speech and singing registers). Thus the impedance of the tube may be designated as

$$Z_t = \frac{\rho c}{A} \qquad 4.5$$

What this means is that for every infinitesimal section of the acoustic tube, the pressure P and the flow U may be related by 4.4. In circuit terms, the pressure becomes voltage, the volume airflow becomes current, and the impedance of the section is given by 4.5. The same types of equations can be used to describe pressure wave propagation through the tracheal walls, although the impedance is much more complicated (Ishizaka *et al.* 1975).

The lumped-element circuit model of the tracheal wall

This section approximates the tracheal wall with a lumped-element circuit model. It simulates the physical properties of the trachea using electrical circuit analogs. Analysis of the electrical components will yield insight as to the relationship between the subglottal pressure and the velocity of the wall.

To begin, we use the model of a section of the vocal tract shown in Figure 4.2 (derived from Flanagan (1965) and Ishizaka *et al.* (1975)). The subglottal pressure driving force $V_s(t)$ is represented by an AC source and the impedance of the tracheal wall ($Z_t(\omega)$) is represented by a resistance R (damping effects), inductance L (inertance effects), and capacitance C (elasticity) . The current I(t) in the system represents the

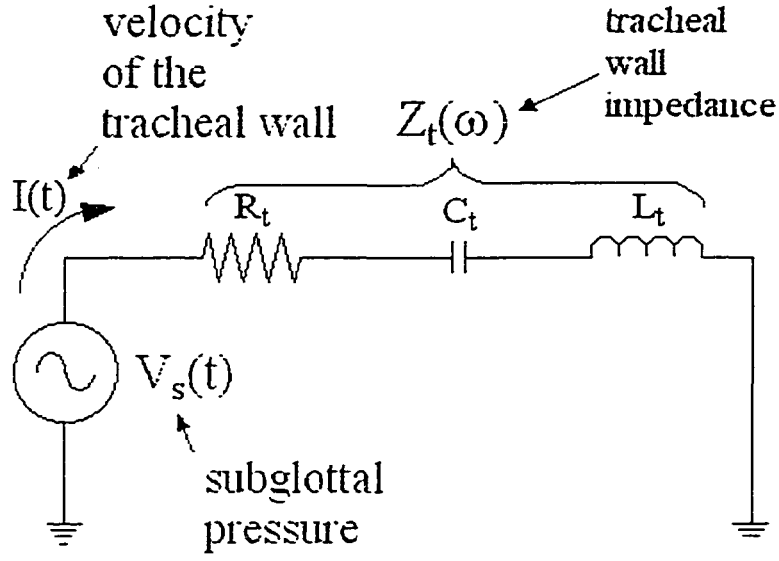


Figure 4.2. Electrical circuit model of the tracheal wall.

"flow" of the tracheal wall - in this situation (an AC system) it is the volume velocity of the tracheal wall. We will differentiate the position signal from the GEMS in order to estimate the wall velocity. We will use the differentiator discussed in Section 2.1.2.

The tracheal wall impedance comes from the reaction of the tracheal wall to the changing subglottal pressure inside of it. A massless, perfectly compliant wall would mirror the subglottal pressure variations perfectly, but a real wall has mass, compliance, and thermal loss and thus is capable of storing and absorbing energy. For these reasons, the wall's reaction to the changing subglottal pressure is not direct; it is normally out of phase with the driving force and is not the same magnitude at all frequencies. Thus the wall velocity I(t) is out of phase with the driving pressure $V_s(t)$. The differences in phase and magnitude depend on the properties of $Z_t(\omega)$.

In this model, we are neglecting the radiation impedance of the trachea, which is normally in series with Z_t. The radiation impedance of a pulsating cylinder can be found

in Rscevkin (1963) – but as its greatest impedance was less than 50 times that of the tracheal wall, it is safely ignored.

From experiments with a mechanical "shaker" to determine the tracheal impedance (it drives a surface and measures the required force and resulting velocity of the surface) performed in 1974 (Ishizaka, French, and Flanagan, 1975), Milenkovik and Mo (1988) derived the following values for R_t, L_t, and C_t:

$$R_t = \frac{B}{2l\sqrt{\pi A}}$$

$$L_t = \frac{M}{2l\sqrt{\pi A}}$$

$$C_t = \frac{2l\sqrt{\pi A}}{k}$$

where

$$M = 2.4 \ \frac{g}{cm^2}$$

$$k = 4.91 \times 10^5 \ \frac{dyne}{cm^3}$$

$$B = 2320 \ \frac{dyne \cdot sec}{cm^3},$$

A is the cross sectional area, and l as the length of the tracheal section. Using a radius of 0.75 cm to simulate the subglottal region, the following per-unit-length values are obtained:

$$R_t = 370 \ \frac{g}{cm^4 \cdot sec}$$

$$L_t = 0.382 \ \frac{g}{cm^4}$$

$$C_t = 1.28\times10^{-5}\ \frac{cm^4\cdot sec^2}{g}$$

leading to an impedance of

$$|Z_t| = \sqrt{370^2 + \left(0.382\omega - \frac{1}{1.28\times10^{-5}\omega}\right)^2}\ \frac{g}{cm^4\cdot sec} \qquad 4.6$$

with

$$\angle Z_t = \tan^{-1}\left(\frac{0.382\omega - \frac{1}{1.28\times10^{-5}\omega}}{370}\right) \qquad 4.7$$

In Figure 4.3, the frequency response of the model is shown. It is clear that the wall impedance rises rapidly with frequency, and therefore acts as a lowpass filter. The magnitude increases at the rate of about 6 dB/octave (an octave is a doubling in frequency). This magnitude increase is unfortunately too rapid, as we saw in Section 2.1.2 that the GEMS-derived velocity (equivalent to the current of the circuit) spectrum of the author's tracheal wall is already quite high in magnitude, losing only about –4 dB/octave instead of the approximately –12 dB/octave normally assigned to the excitation function. The inductance L of the circuit is the cause of the increase in magnitude response with frequency. The magnitude and phase responses vs. frequency for different values of L are plotted in Figure 4.4. It is clear that the slope of the magnitude increase is substantially less for smaller L. Since the inductance is proportional to the mass of the wall, and the measurements were taken on the outside of the neck, which includes the skin, fat, fascia, muscle tissue, and cartilage (see Figures

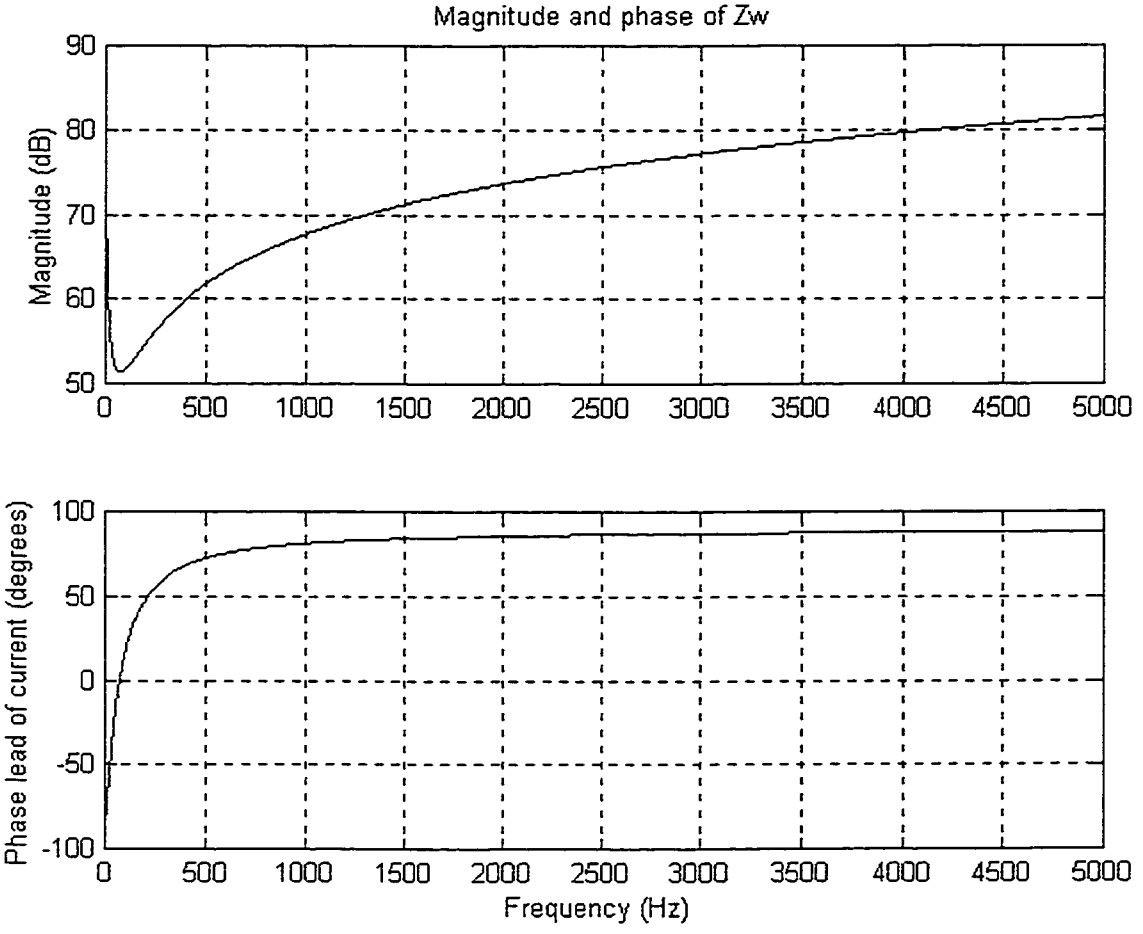


Figure 4.3. Magnitude and phase lead of the impedance Z_w of the lumped-element circuit model of the vocal tract.

3.19 and 3.20); it is certainly plausible that the calculated value for L could be off by a factor of 10.

Therefore, in order to reduce the high frequency noise in the calculated pressure signal, in the next section we will use a value of L/10 for the inductance of the system. The other circuit values remain unchanged, although it is quite likely that the capacitance C is undervalued, as the compliance of the rear wall is probably much greater than that of the frontal trachea with its supporting cartilage. Raising C does not change the response dramatically, as seen in Figure 4.5. In the future, though, it would prove instructive to

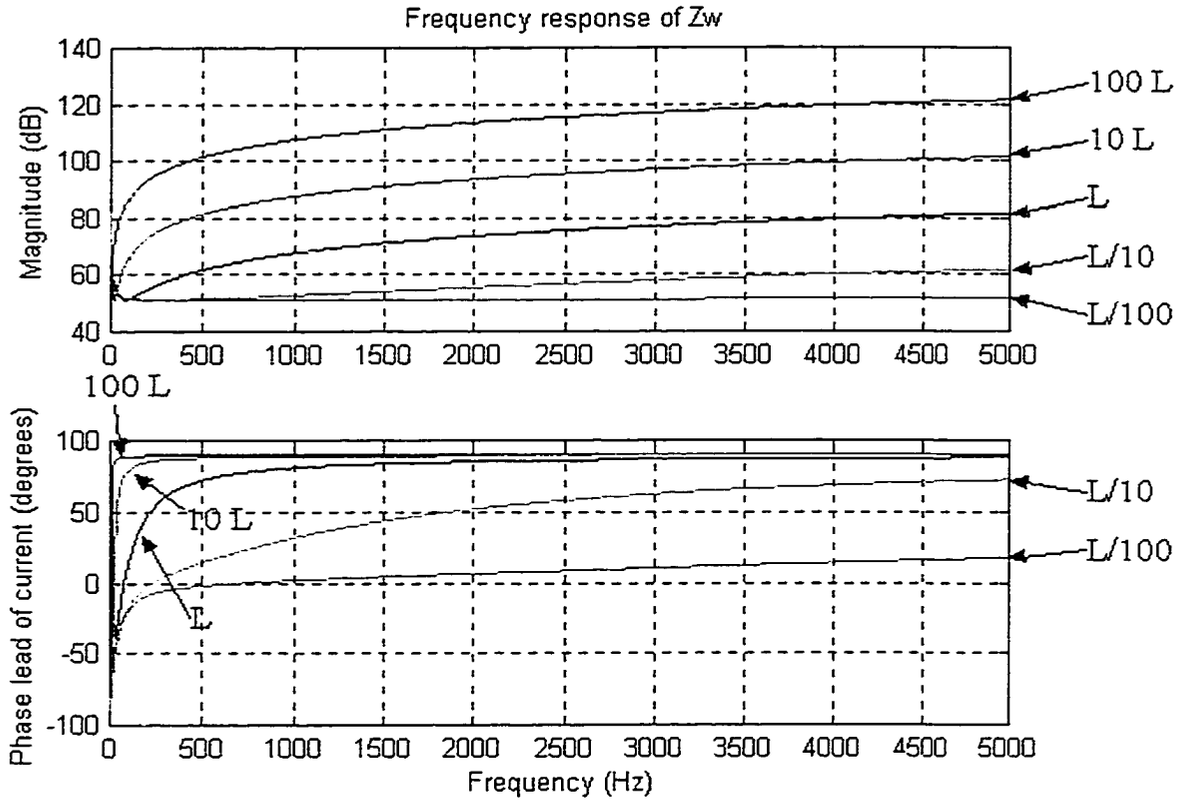


Figure 4.4. Plot of modeled tracheal wall frequency response vs. L.

take measurements of the posterior tracheal wall directly using cadaver tissue to better simulate its properties.

Modeling the analog model digitally

We must now model the tracheal wall impedance digitally so that we can use it to filter the velocity signal from the GEMS. Since the subglottal pressure (voltage) is equal to the velocity of the wall (current) multiplied by the complex tracheal wall impedance, once we have a satisfactory model we may simply filter the velocity signal with the model to produce the driving subglottal pressure.

I constructed an appropriate digital model of the tracheal wall using the method discussed in Section 2.1.2, where a suitable magnitude fit is derived and then the phase is

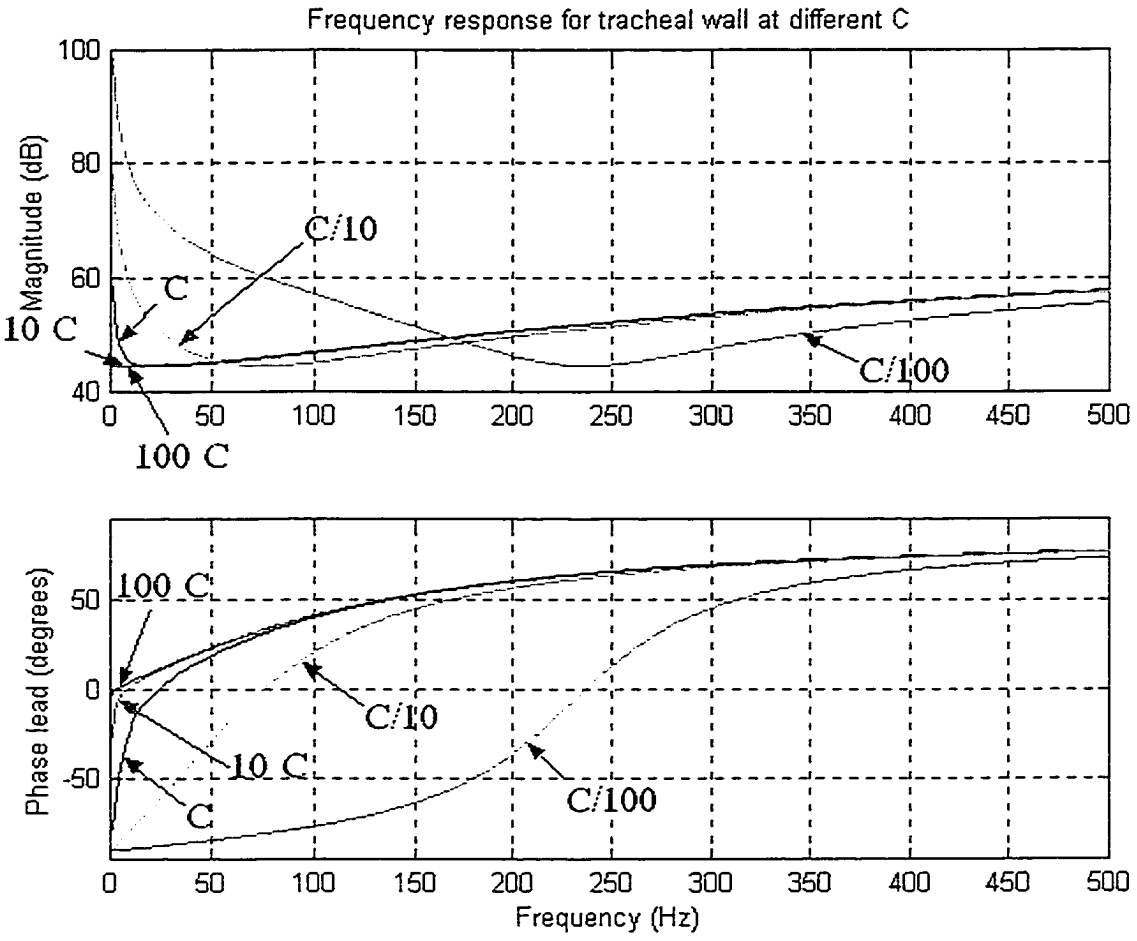


Figure 4.5. Plot of modeled tracheal wall frequency response vs. C.

corrected using anticausal allpass filters. The results are shown by the blue (circuit response), red (modeled response), and black (model with allpass) traces in Figure 4.6.

The magnitude fit (using a fourth order digital filter created using the "yulewalk" command in Matlab) is quite good except for frequencies less than about 75 Hz. The phase fit of the model (bottom plot, red line) is quite poor. This would result in substantial distortion of the wave if not corrected. Using the anti-causal allpass filter techniques described in Appendix A, I was able to correct the phase to that seen in the black line of the bottom plot of Figure 4.6 using three anti-causal allpass filters. For the frequencies above 4500 Hz that could not be corrected, a 15th order Chebychev II (max

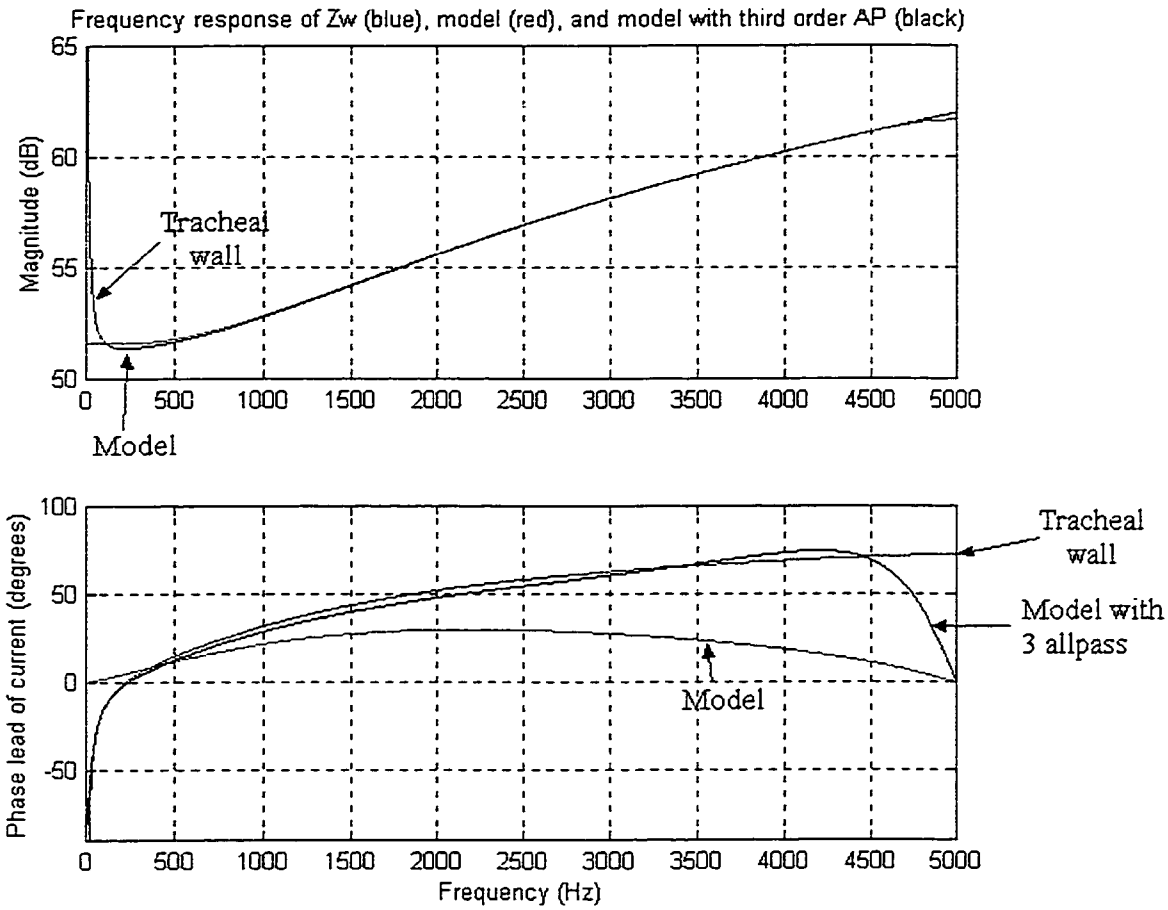


Figure 4.6. Tracheal wall impedance modeled digitally.

flat in the passband) filter was used to remove frequencies greater than about 4 kHz. The phase distortion could not be completely removed as it was in Section 2.1.2 as the tracheal wall model is a fourth order model, much more complex than the simple first order model used in Section 2.1.2.

Now we are ready to determine the driving subglottal pressure. The steps we follow are:

1) Inverse filter the raw GEMS signal to recover the posterior wall position vs. time
2) Differentiate the position signal using the noncausal differentiator discussed in Section 2.1.2 to get the wall velocity signal.

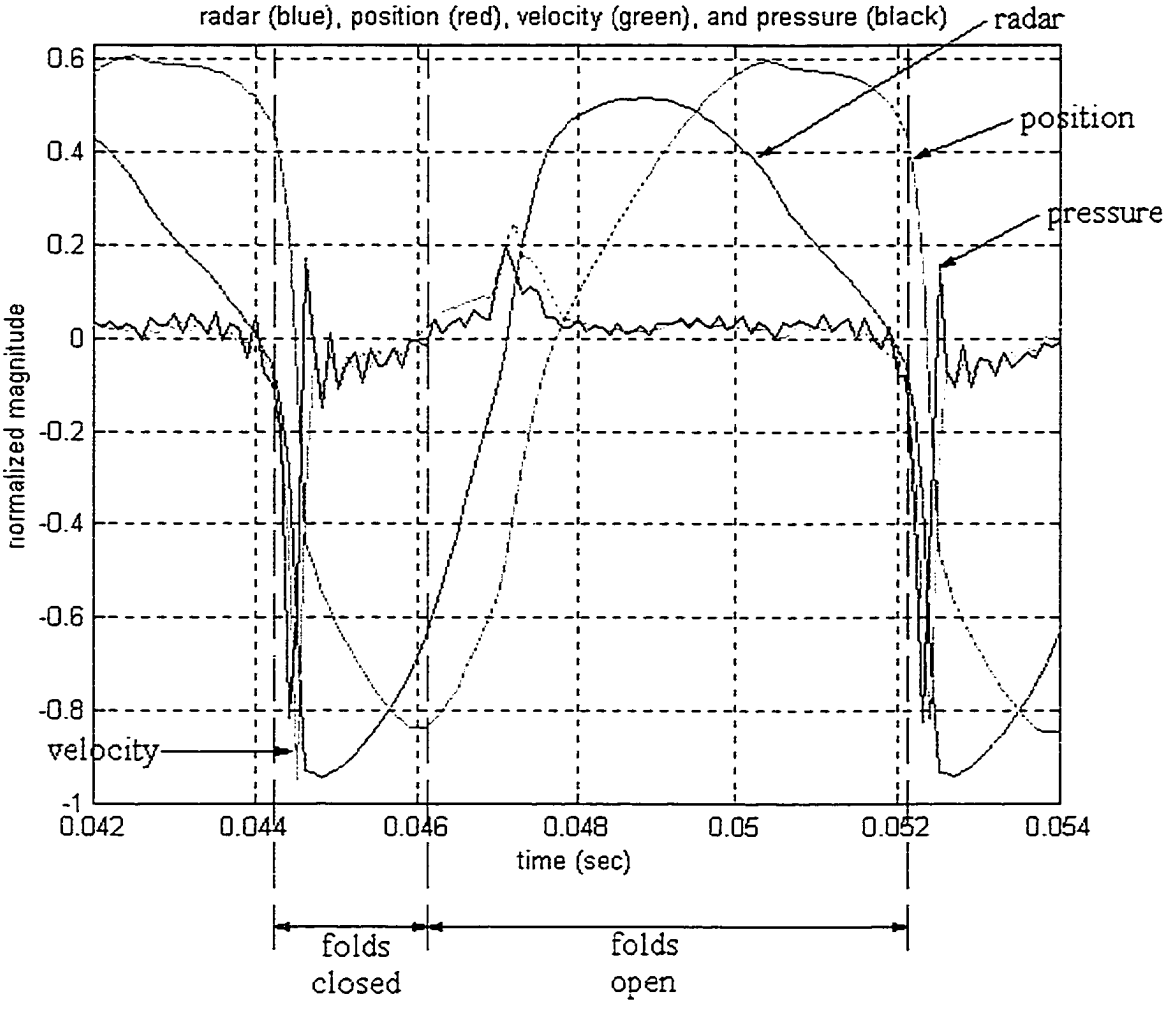


Figure 4.7. GEMS, position, velocity, and pressure for subject GB.

3) Filter the velocity signal using the model above to get the driving subglottal pressure.

The results are shown in figures 4.7 and 4.8. In Figure 4.7, the GEMS, position, velocity, and pressure signals are shown, whereas in Figure 4.8 the original GEMS signal and the inverted pressure are plotted alone to facilitate comparison. All the examples are from a single individual (the author), but other subjects' signals looked similar.

Discussion on the position, velocity, and pressure

In this section, we will ask the following question: Do the tracheal wall position, velocity, and pressure signals make sense when compared to what we know about the fold motion?

The position certainly does make sense from what we know about the fold operation with respect to the GEMS signal (see Section 3.3). The position signal is at a maximum just before the folds begin to close. Using the sensitivity envelope in Section 2.1.3 and the null positions in Section 3.6, we determined that the sensitivity is positive at the distance from the GEMS to the rear wall (about 85 cm), so this denotes the position closest to the GEMS. At this time the posterior tracheal wall has been pulled inward by the Bernoulli force of the rushing air and the relaxation of the elastic wall, and is about to halt its motion and begin to expand out as the folds begin to close and the glottal resistance rises. Soon after, the folds close completely and the trachea wall expands rapidly, moving away from the GEMS. At this time, the GEMS indicates a rapid change in position away from the GEMS. So, the wall position signal makes sense in terms of polarity and magnitude from our knowledge of the operation of the folds. Since the velocity is just the derivative of the position, it corresponds with the posterior tracheal wall motion as well.

The pressure (more easily seen in Figure 4.8) also seems to coincide with our knowledge of how the pressure inside the trachea varies. Soon after closure occurs, the GEMS-derived pressure shows a large negative pressure spike due to the sudden cutoff of flow. Since the subglottal pressure is actually positive at this point, we may conclude that we are displaying our derived pressure signal with the wrong polarity. Recall from Chapter 2 that the GEMS signal is ambiguous in sign, which we rectified by determining

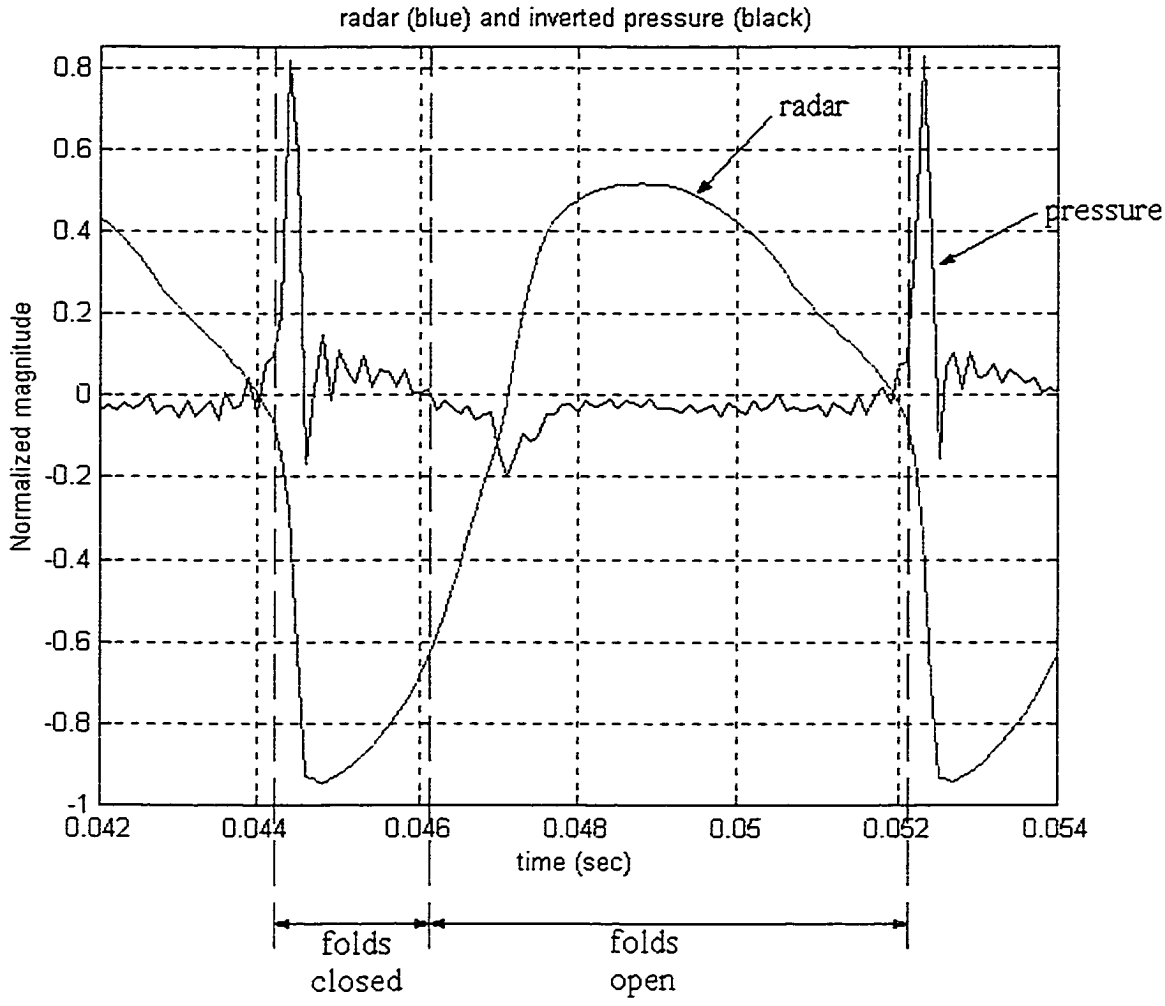


Figure 4.8. GEMS and inverted derived pressure from subject GB.

the sensitivity curve. The sensitivity was defined as positive for the distance that the rear tracheal wall is located, but this was an arbitrary decision we made during the shaker experiments. Thus the polarity of the calculated pressure signal is simply inverted to match the actual pressure change, and it is graphed inverted in Figure 4.8 in order to facilitate the following discussion.

Inspecting the GEMS-derived pressure signal in Figure 4.8, we see that after the large positive spike (indicating a sudden increase in pressure) a small negative spike follows, indicating a sudden decrease in pressure. This is most likely due to a rarefaction

that occurs once the high-pressure compression signal rebounds from the folds and travels toward the lungs. The pressure quickly becomes positive again and stays there until the first small openings in the folds occur (denoted in the graph). During this time the tracheal wall continues to balloon outward under the trapped subglottal pressure. As the folds open, some airflow begins and the subglottal pressure drops, but it does not drop substantially until the folds open much more, denoted by the small negative peak in the pressure at about 0.047 seconds. The negative peak is much smaller in magnitude as the opening of the folds is much more gradual than the closure. The pressure stays slightly negative as flow is occurring, and only becomes positive again immediately before closure. Thus, the general behavior of the GEMS-derived pressure signal corresponds with the changes in pressure that occur in the subglottal region of the tract.

The similarity noted between the breathy audio and the GEMS in Section 3.4

As promised, we will now examine the breathy audio and the GEMS-derived pressure from Figure 3.14. A plot of the data is shown in Figure 4.9. It is clear that the phase of the audio and GEMS-derived pressure is very similar, although the pressure seems to have much more high frequency noise than the audio. This can be interpreted as a sign that we are on the right track, but still have some progress to make on the pressure transformation.

Determining the excitation function

The subglottal pressure (or equivalently from the 1-D model described in equation 4.4, the subglottal airflow) can be considered to be the excitation function when the folds are open and if it is measured near the glottis, as we do. When the folds are open, the calculated subglottal pressure is a good estimate of the pressure wave injected into the

vocal tract, and thus may be considered to be an excitation function for the vocal tract. When the folds are closed, there is no excitation source for the vocal tract, as the vocal tract is a closed pipe at the glottis. There is no new energy injected into the tract at this time, although there are reflections taking place in the tract from the pressure pulse caused by the closing folds. However, the rarefaction pressure wave created supraglottaly when the folds close should be somewhat similar to the compression pressure wave created subglottally, as the same mechanism (the rapid interruption of airflow) is responsible for their creation. Waves in the two regions would just differ in polarity, which does not change the spectral content of the wave. Indeed, we observed this change in polarity in Section 3.4 (see Figure 3.19). For the remainder of this paper, then, we will consider the GEMS-derived pressure signal calculated above to be our excitation function.

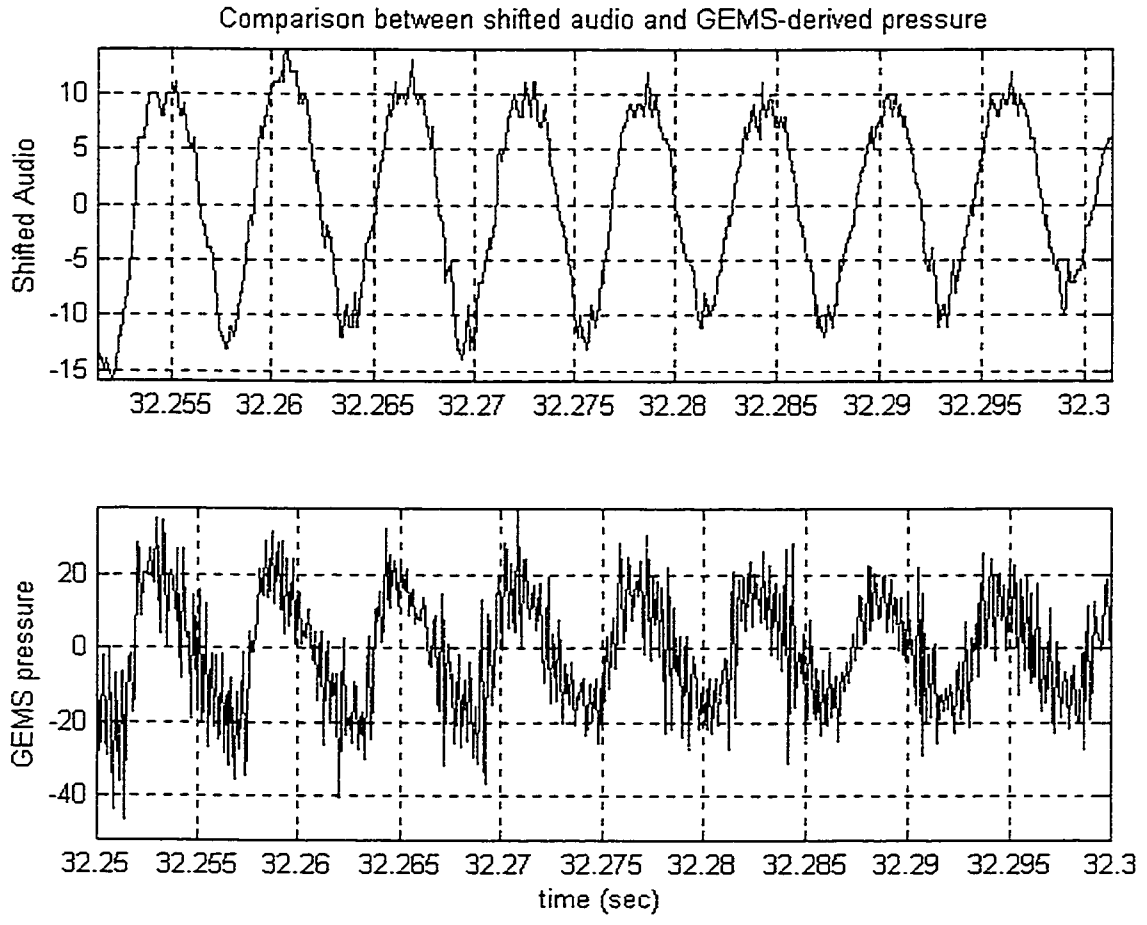


Figure 4.9. The breathy audio and the GEMS-derived pressure from Figure 3.14.

Section 4.2. Human vocal tract transfer function calculations

Now that we have an excitation function for the system, we may calculate the transfer function for the vocal tract. An ARMA model was used to calculate the transfer function for three subjects phonating three vowels: /ɑ/, /i/, and /u/ (see Appendix D for phonetic descriptions and Section 2.4.2 for a review on transfer function models). Matlab 5.1's routine **arx.m** was used, which calculates a transfer function given the input and output of the system, and the number of poles, zeros, and the delay. The input for these calculations was the GEMS-derived pressure, and the output was the audio signal recorded using a condenser microphone. We used 14 poles and zeros (which are covered in Appendix A), and a delay of 13. The delay specifies the time (in samples of the discrete system) that it takes for an input to be processed by the system and emerge as an output. The travel time from the glottis to the microphone (a distance of about 50 cm) is on the order of 1.3 milliseconds. At a sampling rate of 10 kHz, 1.3 milliseconds is 13 samples.

As the transfer functions are calculated, the formants are calculated by a simple peak-picking algorithm. The first four formant locations are stored for later illustration of the stability of the calculation.

The transfer functions for the phonemes /ɑ/, /i/, and /u/ for the subject GB are plotted in figures 4.10, 4.11, and 4.12 at the end of this section. The locations of the first four formants are given at the bottom of the plot. The first transfer function, /ɑ/, has an extra formant located at 371 Hz, which is possibly due to over-specification of the model (i.e. we are using too many poles to model the function). It might also be that the peak at 371 Hz is an accurate representative of the physical tract. With this exception, the

formant locations match those predicted in Titze (1994, p.149-150), so we may feel confident that our methods are capturing the characteristics of the vocal tract. However, it will take more research into the transfer function calculation methods employed on a variety of individuals to assure that our transfer functions are accurate, stable and repeatable. This study is underway at this time.

The transfer function for /ɑ/ does show a good zero at about 1850 Hz, and a strong pole at 2793 Hz. The transfer function for /i/ has a very flat section between the first two resonances, indicating a strong zero there. The transfer function for /u/ is dominated by the first formant, with following formants relatively evenly spaced.

In most cases (the only exception being the fourth peak in /ɑ/), formants located below about 2500 Hz are much sharper than those above. The broadening of the formants indicates greater energy loss at the higher frequencies (Titze 1994, pp. 146-148). The transfer functions are also stable, with the magnitude decreasing and finite at both 0 Hz and 5000 Hz.

The next three plots (Figures 4.13 to 4.15) diagram the relative stability of the formant locations for subjects GB and TG for another set of the three phonemes above over an 800 msec sample length. The formant locations are relatively stable, with occasional imperfections, but the results are encouraging at this early stage. Also, note the similarities and differences in the locations of the formants for the same phoneme but for different individuals.

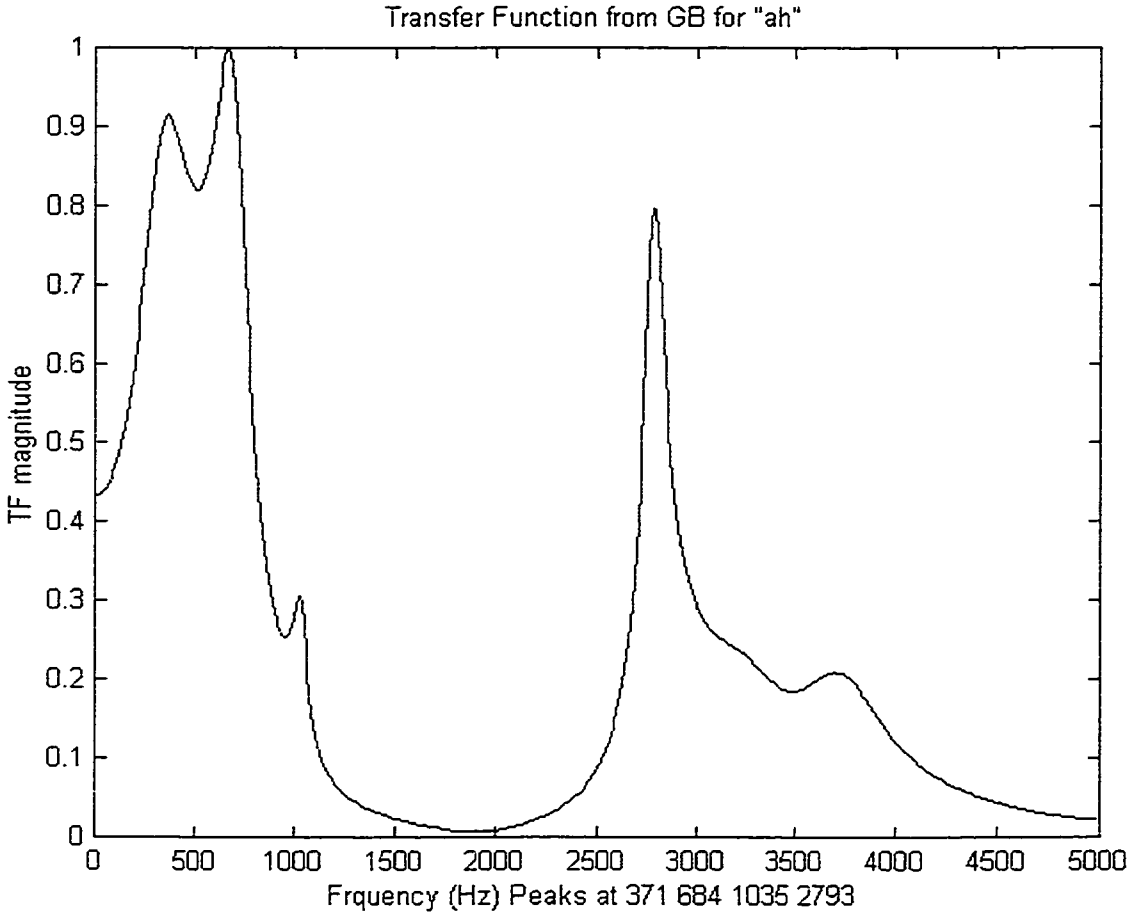


Figure 4.10. Calculated transfer function for /a/, subject GB. The first formant has been split into two formants, possibly due to using too many poles to model the transfer function. The average of the first two formant locations is 527 Hz, and the third is at 1162 Hz. Normal locations for the first two formants are 600-1300 and 1000-1500.

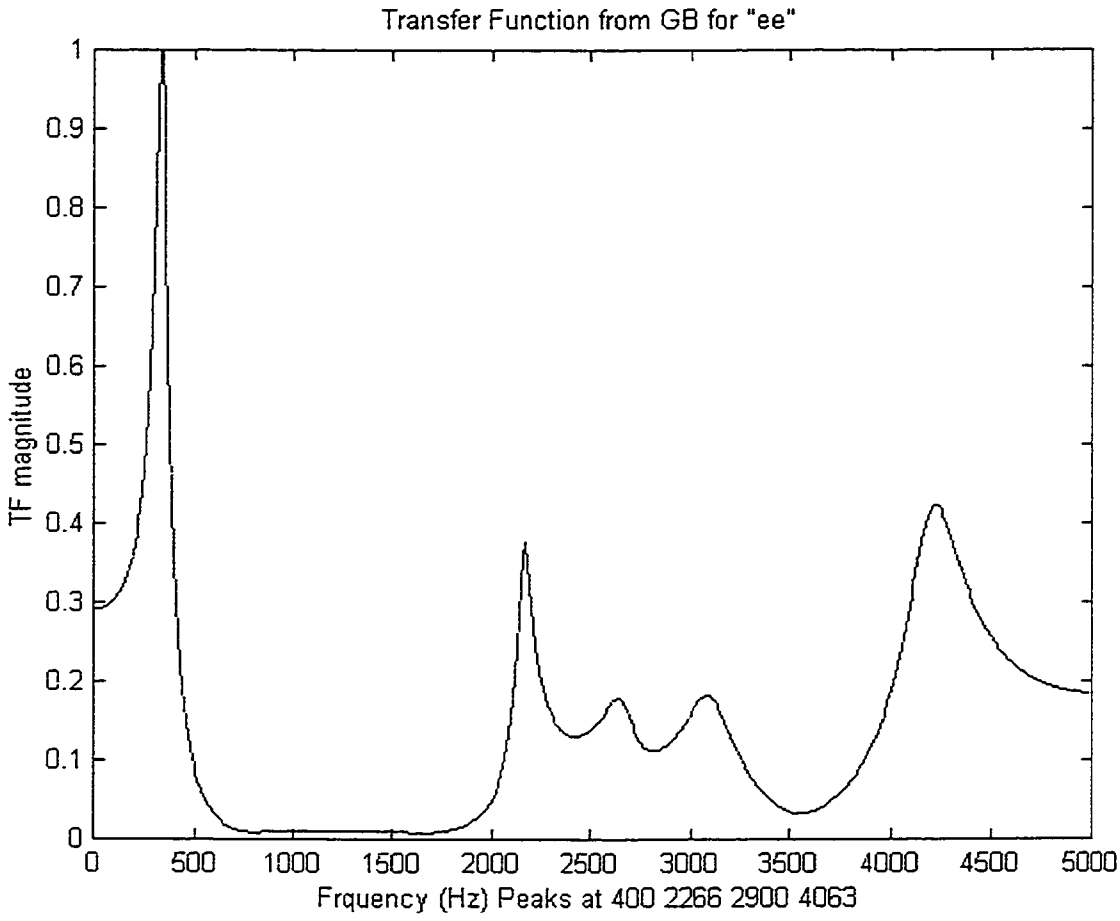


Figure 4.11. Calculated transfer function for /i/, subject GB. The first two formant locations are at 400 and 2266 Hz, normal locations are 200-400 and 2000-4000.

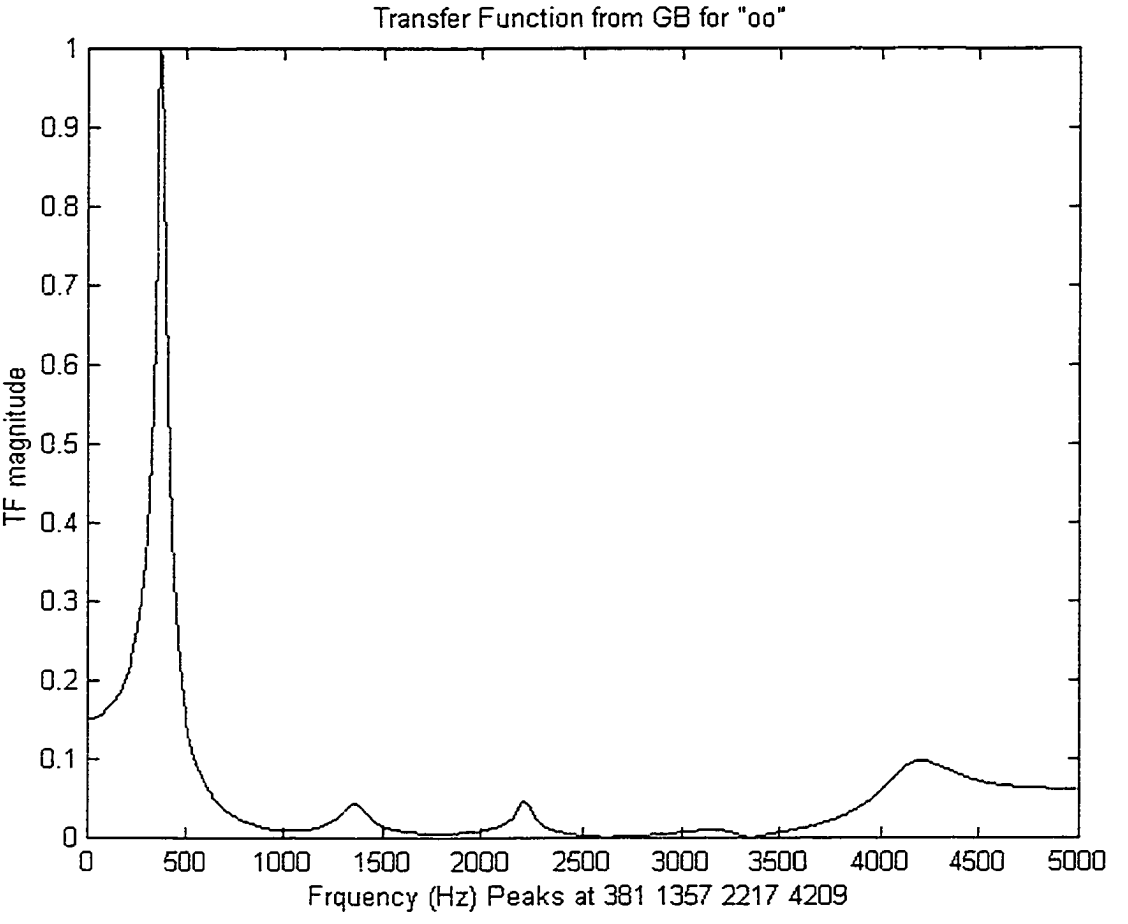


Figure 4.12. Calculated transfer function for /u/, subject GB. The first two formant locations are at 381 and 1357 Hz, normal locations are 400-600 and 900-1400.

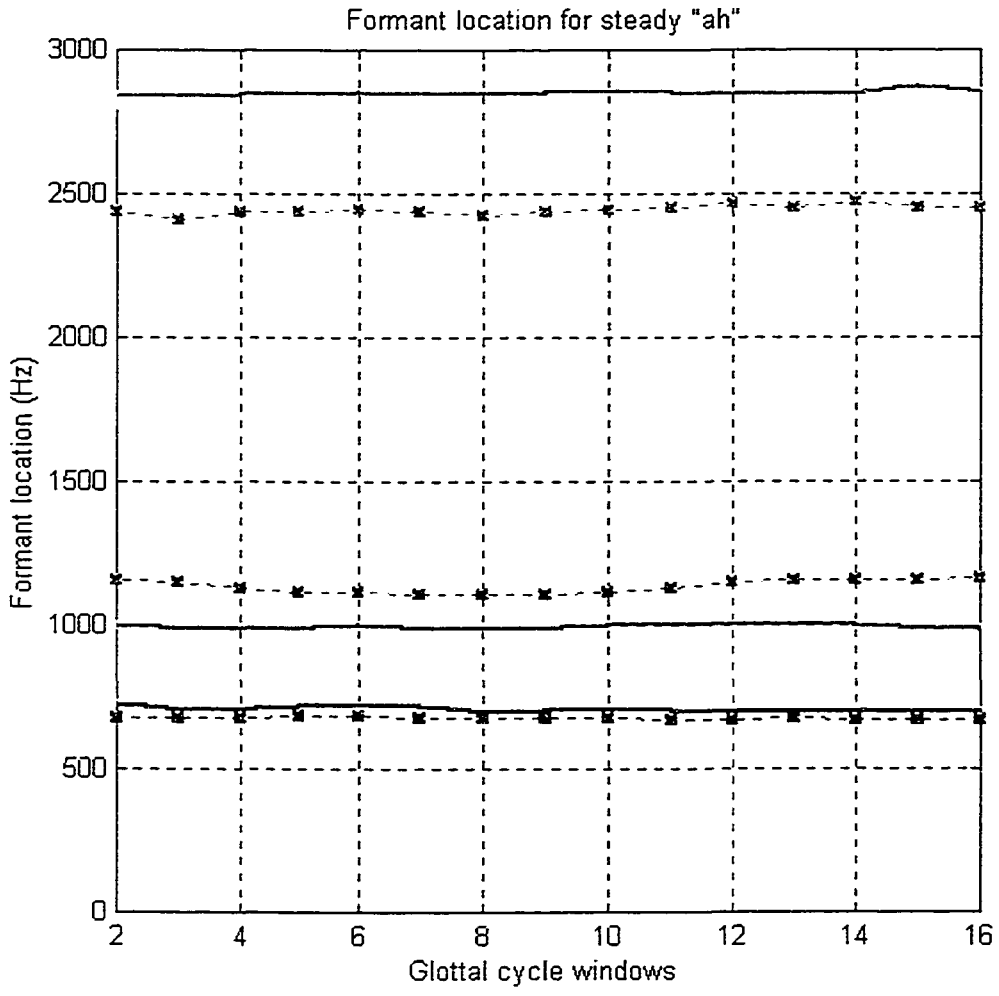


Figure 4.13. Formant location trend for /a/ for subjects GB and TG. Note the relative differences between formant locations, which are individualistic.

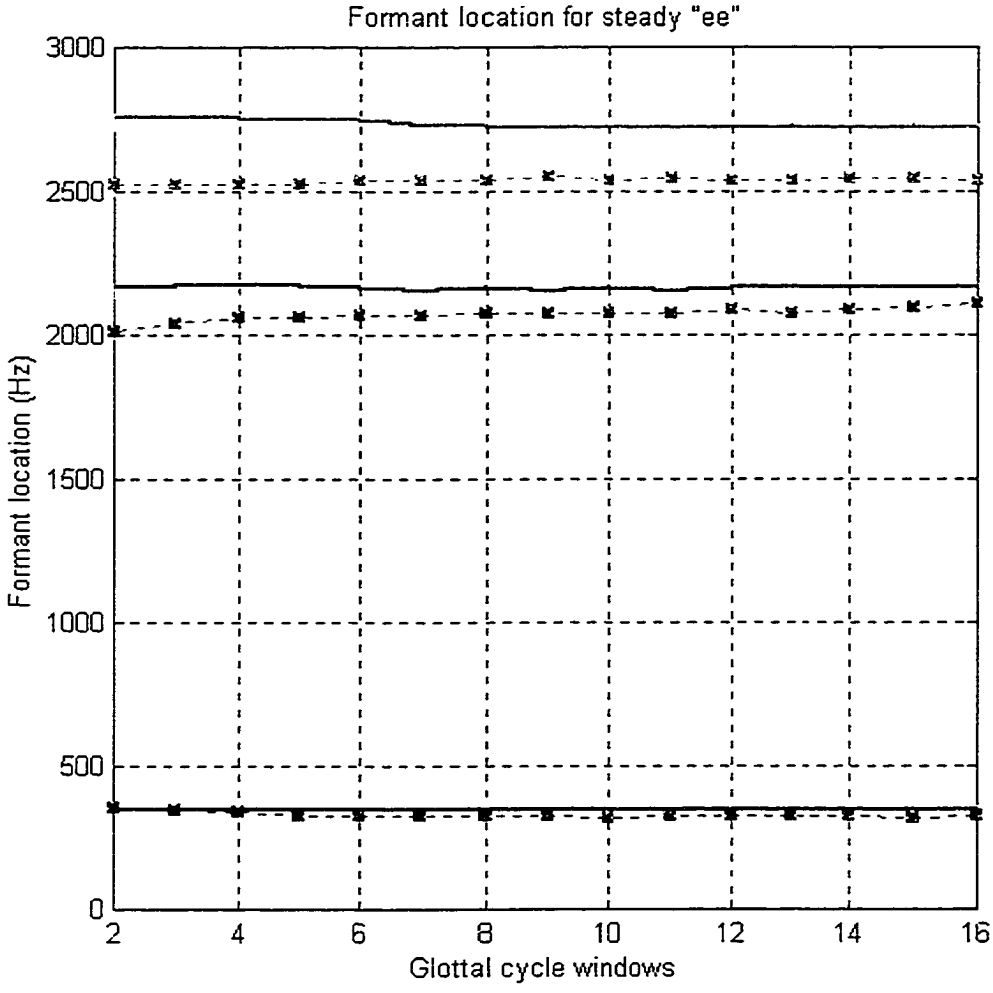


Figure 4.14. Formant location trend for /i/ for subjects GB and TG.

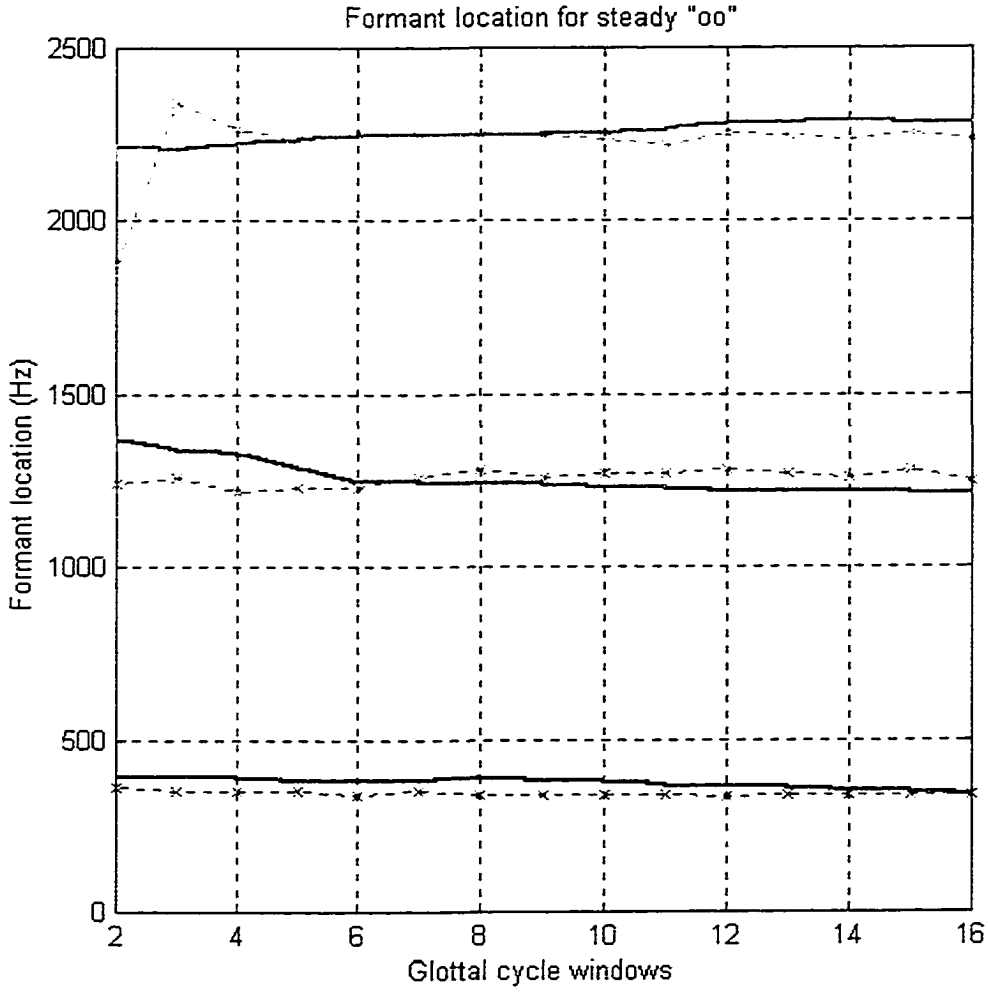


Figure 4.15. Formant location trend for /u/ for subjects GB and TG.

Chapter 5. Conclusion

In this thesis, a new type of sensor has been introduced and quantified. This sensor is known as the Glottal Electromagnetic Micropower Sensor, referred to as either the GEMS. It detects changes in the reflectivity of objects in its field of view. The position of an object moving at a frequency above about 50 Hz can be reproduced quite accurately by removing the distortion caused by the GEMS's analog filters. The magnitude of the GEMS response to sinusoidal motion varies near-sinusoidally in air, with a distance between null points (places of near-zero return regardless of the amplitude of the velocity under observation) of about 35 mm. This is caused by using a homodyne-based detection system.

Several experiments have been described that were undertaken to determine the physiological basis of the GEMS. These experiments included:

1) A GEMS filter characterization experiment, where the GEMS's internal analog filters were characterized, modeled with digital filters, and their distortion removed using a noncausal combination of highpass and allpass filters. This provided us with an accurate representation of the reflection interface's position vs. time.
2) The shaker experiment, which used an industrial shaker head to quantify the GEMS's response to motion. From this we measured the sensitivity of the sensor vs. distance (the sensitivity envelope) and the location of the sensitivity nulls. This allowed us to calibrate the sensor so that if the effective path length from the GEMS to the interface was known, the direction of motion could be derived.

3) 2-D electromagnetic simulations of the reflectivity of the glottal area, including trachea and vocal fold configurations. This substantiated the hypothesis that the trachea was responsible for the majority of the GEMS signal.
4) The capture and analysis of high-speed digital video images of the vocal folds in operation with the simultaneous recording of audio and GEMS signals. Data from normal and abnormal folds were collected and compared.
5) The collection of simultaneous audio, GEMS, and EGG signal for a variety of registers, intensities, pitch, and positions from four subjects at the University of Iowa.
6) A CT scan of the author's neck and head at the University of California at Davis Medical Center. This data enabled accurate estimation of the thickness of the tissue layers between the skin of the neck and the trachea. Using this information and measurements for the indices of refraction of the tissue layers, the effective path length (EPL) through the tissue for EM wave propagation was calculated.
7) Sensitivity null vs. position experiments, where the distance from the neck to the GEMS when a null in the sensitivity envelope was encountered was recorded. The total EPL from the GEMS at the null positions to both the anterior and posterior wall was calculated using the CT scan information.

These experiments showed conclusively that the physiological basis for the GEMS return was the posterior wall of the trachea. A lumped-element circuit model was then constructed using accepted values for the inertance, compliance, and resistance of the neck modified for the decreased mass of the posterior tracheal wall. The subglottal

pressure vs. time was then used as an approximation to the excitation function of the vocal tract.

Section 5.1. Suggestions for future work

The largest improvement that can be made upon this work is the accurate characterization of the posterior wall of the trachea. If the physical properties of the rear wall can be modeled more accurately than at present, it might result in a more accurate excitation and transfer function.

The second area that needs solidifying is the modeling of the vocal tract using a finite number of poles and zeros. It is not clear at this time how this should be done so that a wide variety of transfer functions can be accommodated. What we are using now (14 poles and 14 zeros) works reasonably well over a large number of phonemes, but it would be nice to really describe each one accurately. Perhaps an adaptive algorithm?

The amount of speech processing work that can be accomplished at this point is immense. The GEMS needs to be tested to see if its ability to deliver accurate pitch, voicing onset and offset, and data parsing along glottal cycles (to name just a few, see the next section) will really prove useful to the speech community. I know the GEMS will prove useful, it's just a matter of getting it out to people and letting them go to work. I hope it happens soon.

Section 5.2. Possible applications of the GEMS signal and the excitation function

Now that we have demonstrated the ability to generate excitation functions for the vocal tract using the GEMS signal, we discuss the possible applications in which the GEMS might prove useful in the sciences of speech and audio processing.

1) Speech processor on/off switch

As the GEMS detects the motion of the trachea due to the changing subglottal pressure, which changes only because the folds are operating and modulating air, at its very simplest the GEMS can detect when a person is using their vocal folds. This detection is unaffected by acoustic noise and can tell a speech processor (such as one in a cell phone or speech recognition program) when the person is or is not speaking. It is far more reliable than a simple squelch control and could markedly extend battery life, as cell phones would not have to process audio signals that do not contain speech.

2) Data parsing, accurate and low cost pitch information

The GEMS signal has a sharp drop through zero at the time when the folds are observed to close. As this is defined as the beginning of the voiced glottal cycle, the GEMS signal may be easily used to detect the beginning of the cycle. This has many advantages.

The first is for data parsing (or windowing), in which the data is cut into sections for processing. Current technology uses blind windows, which are the same length regardless of the type or pitch of speech. Windows of 30-40 milliseconds are common, with a 10-20 millisecond overlap between windows. Such intensive processing is needed because the placement of the glottal cycle is not known.

With the GEMS, data may be cut into windows consisting of n glottal cycles, with n any integer. We know exactly where each cycle begins and ends and so we may use adaptive windowing, where the window is sized according to the pitch of the cycles under study. We have found it most useful to use windows that are 2 glottal cycles long, as there is not much change in the vocal tract over this time period (10-20 msec) and 10-20 msec allows for adequate resolution in our Fourier transforms. In addition, the folds may start and stop vibrating in only 2-3 cycles, so it allows us to detect speech onset and offset reliably. As a bonus, Fourier transforms are more accurate and stable when the signal under transformation contains a whole number of wavelengths of the fundamental frequency.

Another application that could utilize the glottal cycle location information is the measurement of pitch. Appendix C contains a paper under consideration for publication by the IEEE Speech and Audio Processing journal. It is entitled "Accurate and noise-robust pitch extraction using low power electromagnetic sensors" and contains many examples of how the GEMS-derived pitch is far superior to any other process in use today. In essence, the GEMS-derived pitch is immune to acoustic noise, far more accurate, and demands about 1% of the processing power required for conventional speech recognition. As the pitch may be measured very accurately as often as every glottal cycle, very small changes in pitch may be distinguished. This would be useful in measuring vibrato (nominally a < 3% change in amplitude at a frequency of 4-6 Hz, Titze 1994) as well as detecting voice disorders that manifest themselves as minute changes in pitch.

3) Transfer functions

We may use the transfer function calculated with the excitation function in many ways to improve speech and audio processing and coding. One of the simplest would be the calculation of formants from the transfer function. The location of the first 2 or 3 formants would specify what phoneme was being voiced during the window (see Figure 2.20 for a graphic of 10 vowels and their first two formant locations, and see the previous section for examples of transfer functions and formant locations).

The transfer functions may be used in many different ways. One of them, speech recognition, could be enhanced by the phoneme detection described above. The phoneme information could be used on its own to build a recognizer, or it could be used to supplement present designs. For phoneme detection, only a small number of poles and zeros are necessary to capture the gross structure of the transfer function. For the applications below, it may be necessary to use more poles and zeros in order to capture the nuances of the transfer function structure.

The calculated transfer functions could also be used for security applications such as speaker verification or identification. Speaker verification verifies the identity of a subject through the parameters associated with his or her voice. Speaker identification identifies a speaker through speech parameters without being told who the subject is. Both methods are currently susceptible to "spoofing" through the use of high quality audio recorders. Since the GEMS measures physical tissue vibration, it is immune to spoofing in such a manner and can be considered a biometric (body-measuring) sensor. It might be that the difference in GEMS signal between individuals could be used as a tool for identification. It is also possible to compare transfer functions for phonemes or words to establish identity. How well this might work is dependent both upon the

repeatability of the GEMS measurements and the consistency of an individual's return. These factors are being evaluated now by a colleague, Todd J. Gable, as he prepares for his thesis on the topic of speaker verification using the GEMS.

The next possible application is that of speaker synthesis. I have purposely used the word "speaker" instead of speech, as we are able to capture the details of a speaker's transfer functions by the use of a sufficient (around 14) number of poles and zeros. We may then reproduce the speech by filtering the excitation function with the transfer function. A library of phonetic transfer functions for an individual may be composed and used with a phonetic speech synthesizer to reproduce an individual's speech. The accuracy of the reproduction would depend only on the accuracy and stability of the transfer functions.

It a similar manner, it appears possible to synthesize musical instruments. The excitation function would be calculated from the position vs. time of a suitable object (such as the string of a violin) and the transfer function of the instrument calculated.

Finally, a possible field of applications for the excitation/transfer function information is voice coders (vocoders). These are the devices and algorithms that digitize speech for digital transmission. The GEMS-derived excitation would be parameterized and transmitted. As it does not change for the same register and intensity of voice, it would need to be updated infrequently. Coded speech segments would require transmission of only the pitch and the transfer function parameters for recombination and synthesis at the receiver.

Appendix A. Inverting a stable filter that is not minimum phase

Statement of problem

Oftentimes a signal processor is presented with a situation that is modeled (appropriately or not) by an unstable digital filter. The challenge is then to take that unstable filter and modify it to get a stable filter without disrupting either the magnitude or phase response or both. This can be difficult if not impossible to do and retain causality, so many times the resulting filter is noncausal but a stable and excellent approximation to the unstable filter. Noncausality is a small price to pay if magnitude and phase information are critical to performance.

Example of a phase-critical situation

In my case it was necessary to build a model of the GEMS's filter characteristics, as the lower 3 dB frequency was low enough (about 70 Hz) to result in noticeable distortion of signal components below about 300 Hz. As a significant amount of energy is contained in the GEMS signal below this frequency (especially for males), it is important to know what the "raw" (undistorted by filters) GEMS signal was so that we could correlate it to the physical position of reflecting interfaces. I therefore used a sweeping frequency generator to inject artificial signals into the GEMS's receiver and from that generated a magnitude and phase versus frequency plot. From that, I built a digital model of the GEMS's filters with the System Identification Toolbox in Matlab using a method described in section 2.1.2. Figure 2.8 shows the resulting plots. Notice that after about 300 Hz, there is not much further change in phase and the magnitude is pretty flat after

about 300 Hz. What we want to do, then, is build a filter that is the perfect inverse of the filters in the GEMS so we may remove their effects from the signal.

Unfortunately, the best fit using the output error model was perfectly stable but not minimum phase, which meant that upon inversion there was a single pole outside the unit circle, denoting instability. Since the filter response has to be inverted so that the effects of the GEMS filters can be removed, we must stabilize this inverse filter. Correcting the phase is important since (for once) we do care what the signal looks like and are not just concerned with its power spectrum.

A brief primer on the z plane

An introduction to the concept of magnitude and phase response in the z plane is warranted to make the following discussion more easily understood.

The poles and zeros of a filter are by definition the roots of the denominator and numerator of the transfer function, respectively. Their values determine everything about the characteristics of the filter – its magnitude, phase response, and stability. The values of the poles and zeros may be plotted on either an x-y (real-imaginary) plot (common in control engineering) or a z-plane plot (more common in signal processing, where z is $re^{j\omega}$, r always positive). Both display the same information in a slightly different format. As I am primarily concerned with the z plane, I will explain the procedure using that as my basis.

As an example I will use a filter with a pole at z = 3/5 and a zero at z = 5/4. This is a rather simple case, but is representative of the type of problem and illustrates the points sufficiently. The case of multiple poles and zeros is easily handled by superposition. The frequency response H(z) is represented by:

$$H(z) = \frac{(z-1.25)}{(z-0.6)} \qquad \text{A.1}$$

This is the form of a highpass filter and is stable, as the only pole is inside the unit circle. The normalized frequency response is shown in figure A.1. Although it is stable, it is not minimum phase, which requires all poles and zeros be located inside the unit circle. When it is inverted, the pole exchanges positions with the zero and is now on the outside, causing instability. The inverse of the filter H(z) is

$$H^{-1}(z) = \frac{(z-0.6)}{(z-1.25)} \qquad \text{A.2}$$

A representative z-plane plot of the inverted, unstable filter $H^{-1}(z)$ is shown in

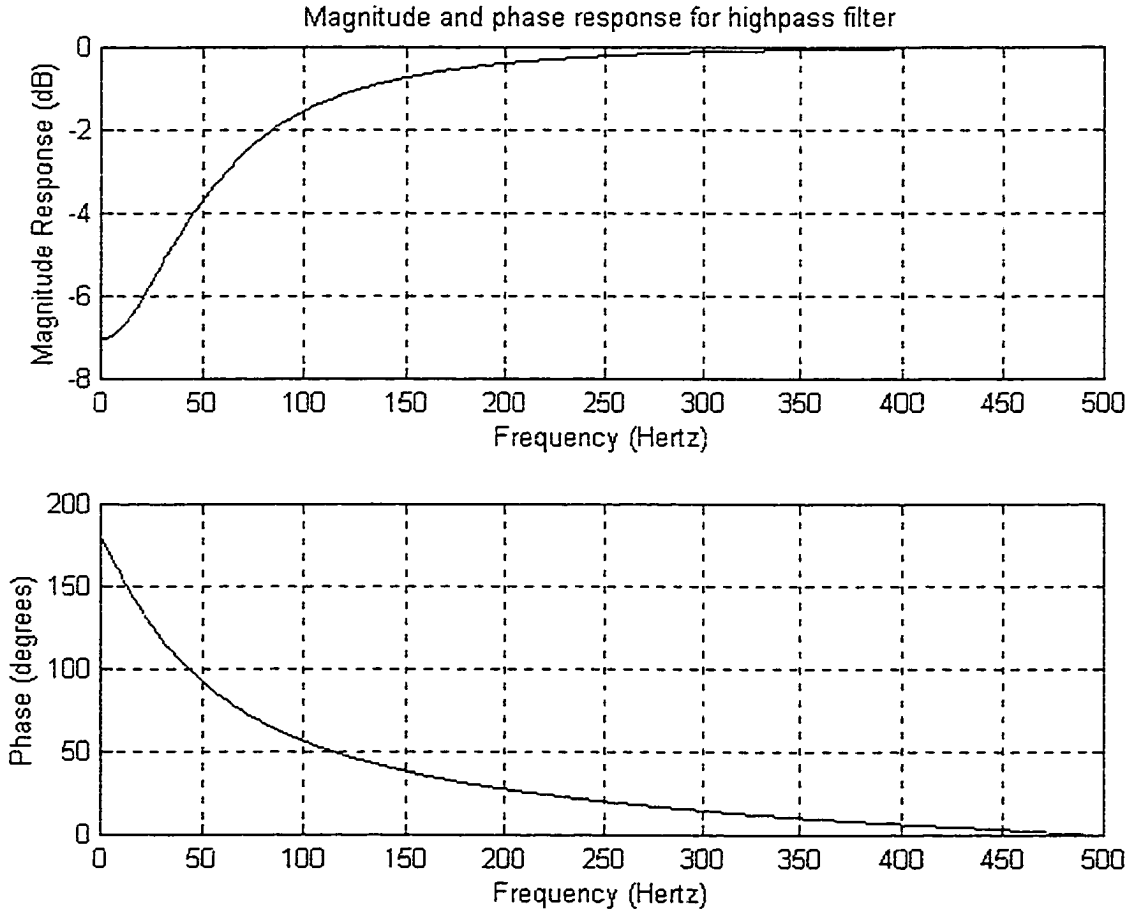


Figure A.1. Normalized frequency response for the example highpass filter.

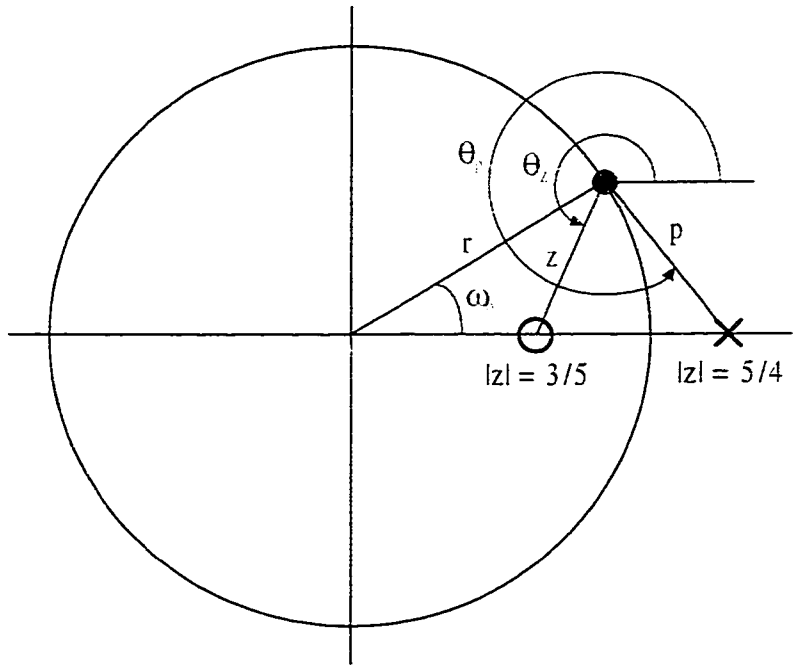


Figure A.2. An unstable filter in the z plane, and the method used to calculate the phase (θ_p and θ_z) and magnitude (z and p) contribution from each pole and zero.

figure A.2.

In the z plane, the unit circle is special in that it represents the points where r = 1, and z reduces to the complex exponential $e^{j\omega}$. On these points, $z = e^{j\omega}$ and the z transform is equal to the Fourier transform. It is from points in the z plane on the unit circle that we determine the frequency and phase response of the system.

At $\omega = 0$, we are on the unit circle and at the point corresponding to DC. As we make our way around the circle, we go through all possible frequencies until we make it to $\omega = \pi$, which corresponds to the highest possible frequency attainable in a digital system, the Nyquist frequency. This is equal to ½ the sampling frequency, and the analog signal sampled by the digital system must have no components equal to or greater than this frequency or the resulting digital signal will be distorted. The lower half of the unit circle adds no new information; it is the mirror of the upper and contains the

response for "negative" frequencies, a mathematical artifact that has little bearing on the real world. Somewhat like an old professor of my acquaintance.

Anyway, determining the magnitude response of the filter at a certain frequency is simply a matter of determining the distance from its position on the unit circle to all the poles and zeros of the system. The magnitude response at ω_0 is given by

$$|H(\omega_0)| = \frac{|z_1| \cdot |z_2| \cdot \ldots |z_n|}{|p_1| \cdot |p_2| \cdot \ldots |p_m|} \qquad \text{A.3}$$

with n the number of zeros and m the number of poles.

Similarly, the phase response of the system is given by

$$\angle H(\omega_0) = \sum_n \theta_n - \sum_m \theta_m. \qquad \text{A.4}$$

Figure A.2 illustrates how these lengths and angles are calculated. From this example it is possible to get an idea how an allpass filter has a flat magnitude response. For an allpass, the zero is located at z = B (which may be complex), and the pole is located at z = 1/B. Thus the distance from any point on the unit circle to the pole/zero pair varies, but the ratio of the distances remains constant.

Now that we have a handle on the z-plane, we can begin the explanation of how we can repair unstable filters using one or more anticausal allpass filters.

Stabilizing unstable filters using allpass noncausal filters

The first step is to build an allpass (AP) filter to cancel out the offending exterior pole while preserving the magnitude response. An allpass filter has by definition unity magnitude response, but the phase is not zero, resulting in a 2-stage filter that now has a different phase response than desired. In Figure A.3, the inverted filter is shown with the pole outside the unit circle at z = 5/4, and the zero inside at z = 3/5. This is an unstable

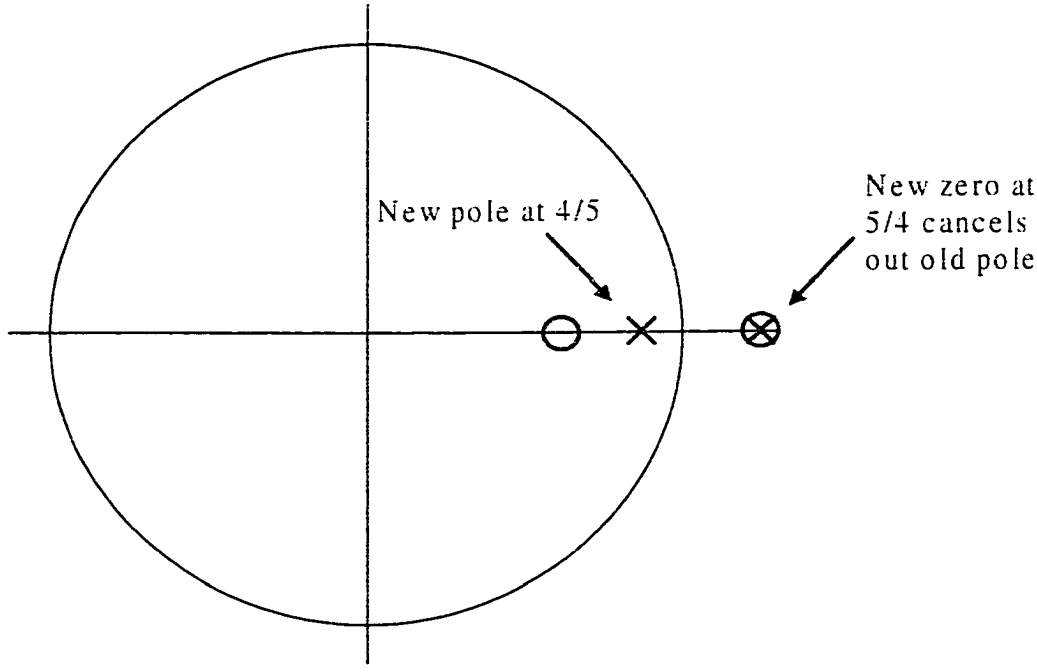


Figure A.3. An unstable filter (black) with a stabilizing AP filter (blue).

filter, and so we build an AP filter with a zero at z = 5/4 and a pole at z = 4/5. The offending pole at 5/4 is cancelled out and a new, minimum phase filter results. As seen in figure A.4, the magnitude response of the filter is inverted, but the phase response is not and is quite distorted. The frequency response of this inverted, stable filter $H_s(z)$ is

$$H_s(z) = \frac{(z-0.6)}{(z-1.25)} \cdot \frac{(z-1.25)}{(z-0.8)} = \frac{(z-0.6)}{(z-0.8)} \qquad \text{A.5}$$

How badly is the phase distorted? Well, it is quite distorted near $\omega = 0$, as the pole is now contributing ~ 0 degrees to the phase where it was contributing ~ π when it was outside the unit circle. It soon becomes clear that in order to cancel out the phase of the original filter we need a filter with negative phase at $\omega = 0$. As you can see in Figure A.2, a negative phase at $\omega = 0$ can only exist if the pole lies outside the unit circle,

rendering the filter unstable. Therefore, in order to get negative phase at $\omega = 0$, we must utilize a noncausal filter.

So now that we have decided on a noncausal filter, how do we go about designing it? It must be an allpass filter so as not to disturb the magnitude response, and the phase should be chosen to be the negative of what we need to fix the distortion. Thus, we design for a certain phase response which, when subtracted from the present phase, will give us something that is closer to the phase response we are trying to get. It may not be perfect, especially for complex filters, but arbitrarily many of these allpass filters may be used so that we can get as close as necessary to the phase we desire.

To begin, we look for a frequency where the phase is important and the distortion is large. In Figure A.4, we see that the maximum error occurs at 0 Hz, but as the lowest voiced frequency in males is normally around 100 Hz, we will try there first. At 100 Hz we determine the correction needed, in this case about 75 degrees or 1.31 radians. As our sampling frequency in this example is 1 kHz, π is equivalent to 500 Hz so that our correction will take place at $\omega = \frac{100}{500} = 0.2\pi$. We need to determine the location of the pole and zero of the allpass filter so that the phase of the allpass is 75 degrees at 100 Hz. In Figure A.5 the new allpass (to be anticausal) filter is plotted along with the inverted, stabilized filter with poor phase response. To the right are three triangles we will use in our calculations. They are magnified and not to scale for clarity.

We have determined that we need 1.31 radians to correct the filter phase at f = 100 Hz ($\omega = 0.2\pi$). From the top triangle in figure A.5, we get the values for x and y:

$$x = \cos(0.2\pi) = 0.809 \qquad \text{A.5}$$

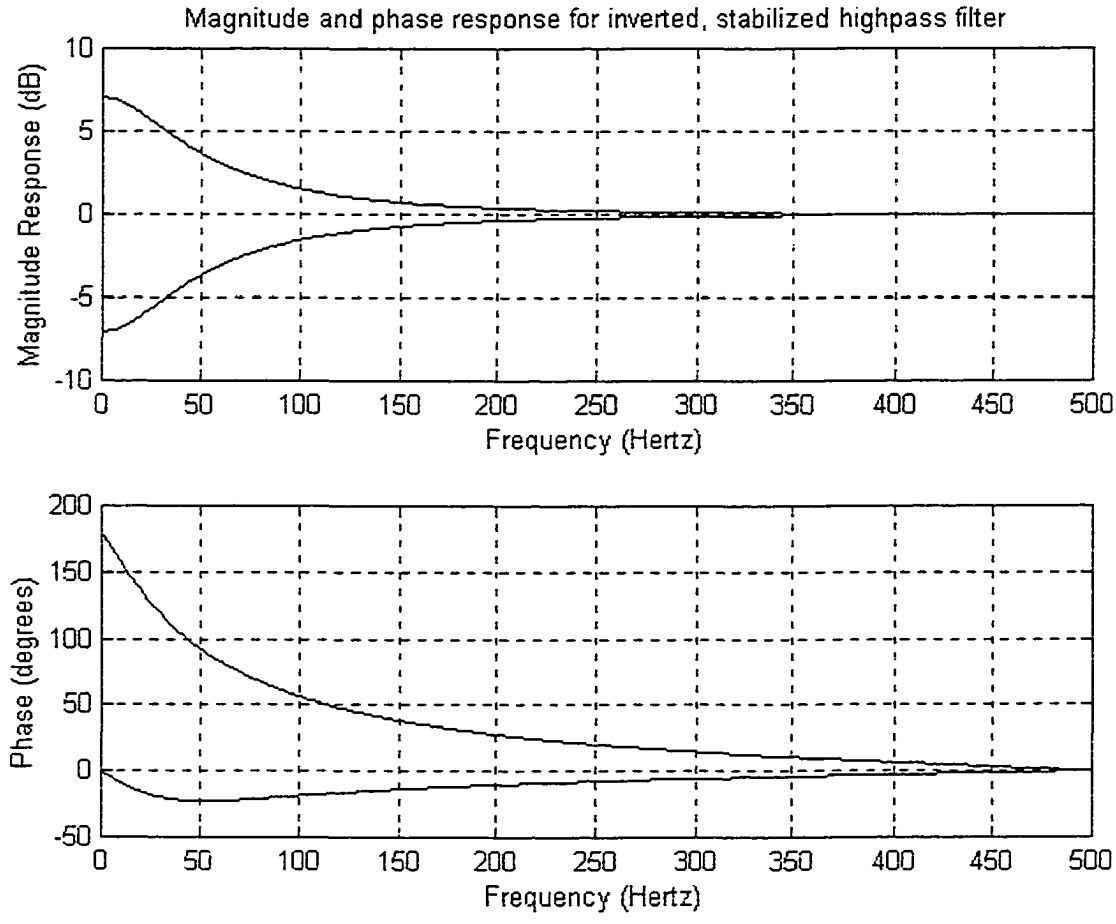


Figure A.4. The inverted, stable filter $H_i(z)$. The magnitude is perfectly inverted but the phase is not.

$$y = \sin(0.2\pi) = 0.588 \qquad \text{A.6}$$

which we can then use to find ϕ_p and ϕ_z:

$$\phi_p = \arctan\left(\frac{y}{x - 1/B}\right) \qquad \text{A.7}$$

$$\phi_z = \arctan\left(\frac{y}{B - x}\right) \qquad \text{A.8}$$

Now, since the phase contributed by the new zero and pole is measured from a horizontal line (parallel to the $\omega = 0$ line, see figure A.2), the total phase of the new AP filter will be

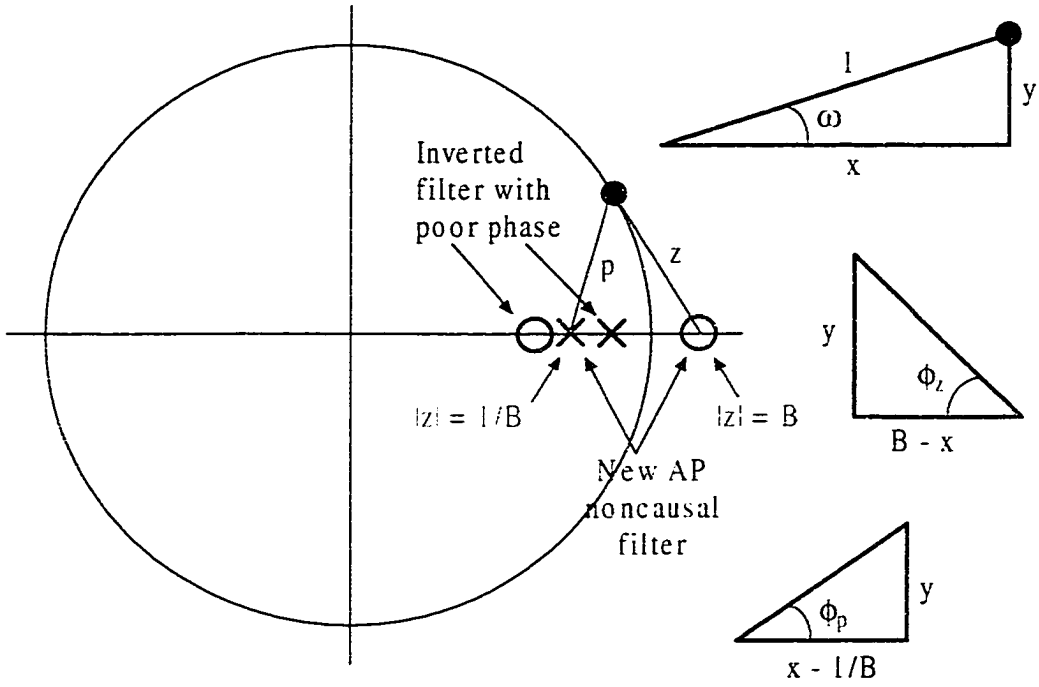


Figure A.5. Plot of second allpass filter (blue) and the triangles used to calculate B, the position of the allpass zero.

$$\Theta = \pi - \phi_p - \phi_z . \qquad \text{A.9}$$

Since we want this to be equal to 1.31, we have the following equation:

$$\pi - \arctan\left(\frac{y}{x - 1/B}\right) - \arctan\left(\frac{y}{B - x}\right) = 1.31 \qquad \text{A.10}$$

or

$$\arctan\left(\frac{y}{x - 1/B}\right) + \arctan\left(\frac{y}{B - x}\right) = 1.83 \qquad \text{A.11}$$

and since

$$\tan(A + B) = \frac{\tan(A) + \tan(B)}{1 - \tan(A)\tan(B)} \qquad \text{A.12}$$

we get

$$\frac{\frac{y}{x-1/B}+\frac{y}{B-x}}{1-\frac{y^2}{(B-x)(x-1/B)}}=-3.75. \qquad \text{A.13}$$

Using the values for x and y in equations A.5 and A.6, we solve for B and find 1.664 to be the only positive answer. This leads us to an AP filter with the transfer function of

$$H_{AP}(z)=\frac{1/B\cdot(1-Bz^{-1})}{(1-1/Bz^{-1})}$$

where the normalization factor (1/B) in the numerator is necessary to make the magnitude equal to one. It comes from the general AP filter form:

$$H_{AP}(z)=\frac{a^*-z^{-1}}{1-az^{-1}}=\frac{a^*(1-z^{-1}/a^*)}{1-az^{-1}} \qquad \text{A.14}$$

where in this case a is real. The phase response for the allpass filter (the magnitude is not shown as it is unity for all frequencies) is shown in Figure A.6. The phase response for 100 Hz is very nearly 75 degrees. It is not exactly 75 degrees due to rounding errors.

The new combined filter $H_c(z)$ is thus

$$H_c(z)=H_s(z)\cdot H_{AP}(z)=\frac{(z-0.6)}{(z-0.8)}\cdot\frac{1/B\cdot(1-Bz^{-1})}{(1-1/Bz^{-1})} \qquad \text{A.15}$$

But we must remember that the addition of the causal allpass filter will make the phase of the combination inverse filter/allpass equal to the original filter at 100 Hz. We want the phase to be the inverse of the original, not equal to it. Therefore we must use the filter backwards in time (anticausally) in order to retain the magnitude response and invert the phase response.

Mathematically this is accomplished by replacing z^{-1} (a delay) with z (an advance), rendering the filter anticausal. To accomplish this in Matlab we just take the coefficients

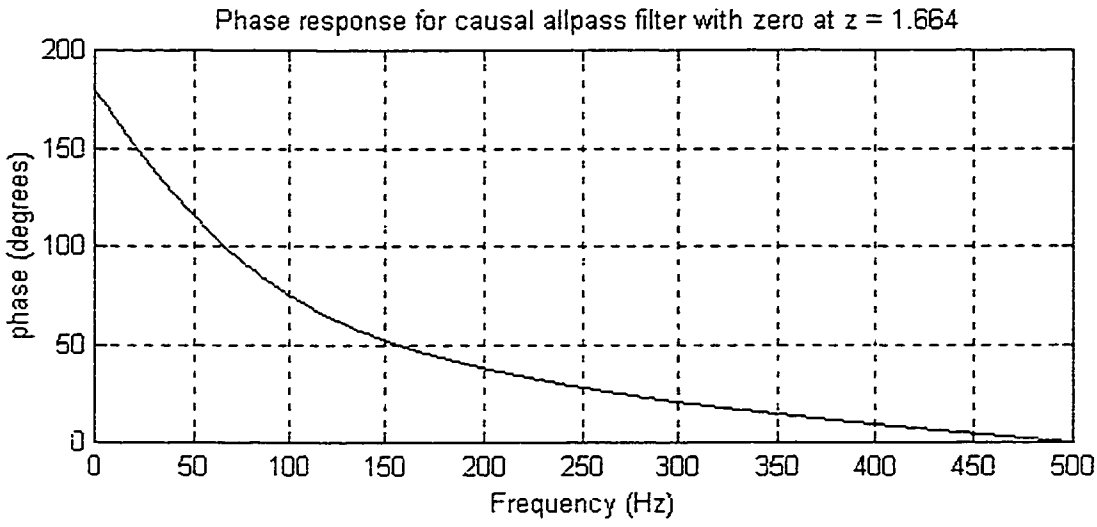


Figure A.6. Phase response for causal allpass filter designed to have a phase shift of 75 degrees at 100 Hz.

of $H_c(z)$ form a filter using the *filter* algorithm. To make it anticausal, we simply take the signal we are going to filter and reverse it in time:

$$x[n] \rightarrow x[N-n] \qquad \text{A.15}$$

where N is the length of the signal. Now we filter the reversed signal with the causal filter to get the phase corrected output:

$$y[n] = \sum_k x[N-n] \cdot h_c[n-k] \qquad \text{A.16}$$

where h_c is the impulse response of $H_c(z)$. The frequency response of the original filter H(z), the stable, inverted filter $H_s(z)$, and the final noncausal filter $H_c(z^{-1})$ are all plotted in Figure A.7. In this simple example, an almost perfect inversion of the phase was possible.

Sadly, for more complex problems this will not be the case. As we are only compensating the phase at a single frequency, at others the phase may still not be a good enough cancellation. Therefore, it is sometimes necessary to try to fit the phase at several

different frequencies near the one you chose to begin with to see if you can get a better fit.

If you have very exact specifications, you may try and add additional causal or anticausal AP to further correct the phase in a particular frequency range, but be warned that the fit near $\omega = 0$ will suffer. This is because placing poles and zeros on the real axis at $\omega = 0$ causes another π to be added to the phase at $\omega = 0$. Thus near $\omega = 0$ the phase correction gets steadily worse as we add AP filters. If, however, your frequencies of interest are not near $\omega = 0$, this technique can work quite well.

Conclusions

After you are finished with the design of your stable, noncausal filter, it is wise to test the result if the inversion is not perfect. This will give you a way to check the distortion still present at different frequencies. A quick way to do this is to build an artificial signal, such as a square wave, filter it using your original model, and then with your inverted and stabilized model. You should get back out the original signal if you have sufficiently compensated for the phase distortion. This is assuming, of course, that your instability resulted from the inversion of a stable filter. If you are trying to compensate for an unstable model, there is no way to compare waveforms directly – you just have to be satisfied that your phase and magnitude plots are accurate enough. We will not perform such a test, as our inversion was perfect and it would not be illustrative.

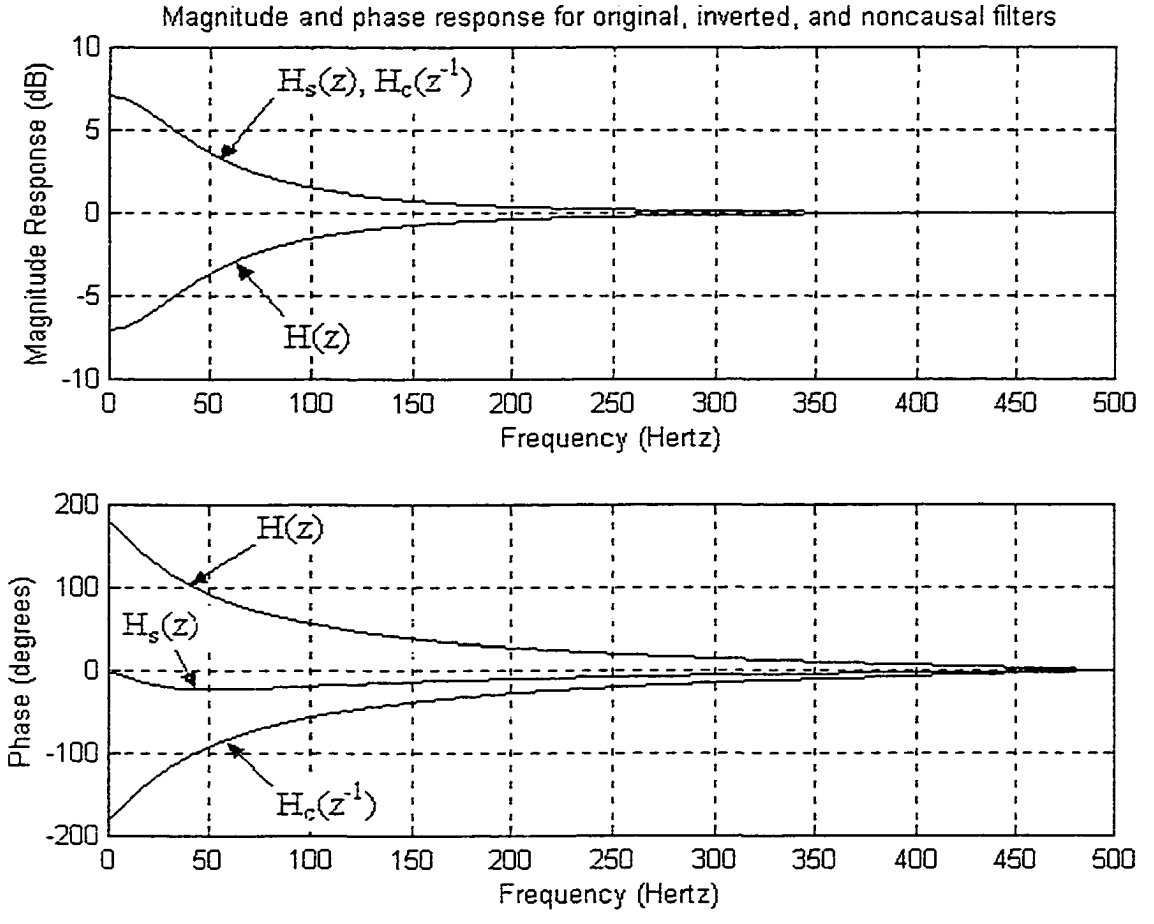


Figure A.7. The frequency response for the original filter H(z), the stable inverted filter $H_s(z)$, and the noncausal filter $H_c(z^{-1})$.

Appendix B.

Use of Kodak EktaPro high-speed digital cameras in laryngoscopy

Gregory C. Burnett & Rebecca Leonard
Lawrence Livermore National Laboratory/UC Davis
UC Davis Medical Center

Abstract: The use of stroboscopic cameras operating at normal speeds (24-30 frames per second or fps) is commonplace in laryngoscopy. However, the time between recorded frames can include several glottal cycles. Thus only disorders which persist for many tens of cycles can be distinguished. In order to observe more transient events, a higher frame rate is needed. High speed (>200 fps) digital cameras available from Kodak provide such an option. These cameras offer good resolution, easy operation, and very fast frame rates (up to 12000 fps) that allow for capture of transitory events such as onsets and offsets of voicing, and details of a single vibratory cycle. These attributes can prove useful to voice clinicians and researchers alike. This paper examines the use of such cameras with and without image intensifiers in a clinical and research setting and offers recommendations for their operation, including the synchronization of the frame information with audio or other data. Specific information is offered on the EktaPro Hi-Spec Motion Analyzer Model 1012 with Intensified Imager Model VSG.

I. Introduction

Kodak has offered high speed digital cameras since 1990, but they are not often used in clinical settings due to their high cost, light requirements and complexity.

Recently, at the UC Davis Medical Center, the authors had the opportunity to use the Kodak EktaPro Model 1012 with model VSG intensifier to validate a new type of electromagnetic sensor developed at the Lawrence Livermore National Laboratory (Holzrichter et al 1998). This sensor, referred to as the GEMS (for Glottal Electromagnetic Micropower Sensor), uses extremely low-power (< 20 milliwatts radiated) electromagnetic waves to detect the ballooning motion of the trachea due to subglottal pressure variations as phonation occurs. In order to more fully understand the physiological basis of the GEMS return, it was necessary to reference the GEMS signal to fold motion to determine if the folds had any direct effect on the GEMS signal. Accordingly, audio, GEMS, supraglottal, and subglottal pressure were recorded along with the external synchronization ("ext sync") signal from the EktaPro intensifier that was used to time-align the signals with the video frames.

A. The EktaPro 1012

The Kodak EktaPro 1012 is capable of capturing 1000 digital fps using the full resolution (239 x 192) or up to 12000 fps using 1/12 of the screen (239 x 16) in 256 shades of gray. There are two separate units that make up the EktaPro – the camera and the processor. The processor is where the processing and timing take place and where the digital frames are stored. The processor can store from 500 to 19659 full frames of data depending on available processor memory. The camera utilizes an NMOS imaging sensor to capture the image and can be joined to the processor using a 15, 50, or 100 foot cable. The camera can accept any lenses that incorporate a standard C-mount. A portable keypad allows the user to direct commands to the processor. The variables that can be controlled via the keypad are frame rate, exposure time (normally 1/frame rate,

overridden by the intensifier controller if installed), portion of the NMOS imaging sensor from which data is collected (from full screen to 1/2, 1/3, 1/4, 1/6, and 1/12), triggering conditions (when and how to start recording), among others.

The EktaPro processor may also be remotely controlled by a computer through a GPIB connection. The digital frame data can be downloaded through this port, but transfer is slow – about one frame a second, or in our case over 40 minutes for a single movie of 2456 frames. Thus the "video out" connection is commonly employed to download the frames onto a SVHS or VHS videotape. Playback is accomplished at 30 fps, resulting in a download time of just over 80 seconds.

B. The Intensifier VSG

For experiments where high frame rates are desired and limited light is available, an intensifier may be used to increase the brightness of the image recorded by the camera. The intensifier option includes an intensified camera and a controller, which sets the gain of the light amplification as well as the exposure time (controlled with an electronic shutter). The increased gain allows frames to be recorded at higher frame rates and/or lower light levels. The shortened exposure times result in sharper images. For the intensifier to be effective, Kodak recommends an exposure time of less than 1/10 of the frame period. Thus for a 1000 fps movie, the exposure time should be less than 100 microseconds.

To aid in signal synchronization, the intensifier controller provides a TTL output (external sync) that is "high" (around 5 Volts) when the exposure is occurring and "low" (around zero Volts) when the (electronic) shutter is closed. This is extremely useful for synchronizing recorded data such as audio with the recorded frames. Together the

shortened exposure times allowed by the intensifier and the ease of synchronization make the intensifier a very desirable addition to the standard EktaPro processor.

C. Frame rate vs. exposure time

Its is important to understand the difference between "frame rate" and "exposure time". The frame rate of a camera describes how many frames it records in a second. The exposure time is the amount of time the NMOS imaging sensor is exposed to light. For the standard EktaPro (without intensifier), the exposure time is normally 1/frame rate, but it can be less. The NMOS imaging sensor is divided into 12 "blocks" of pixels. Each block consists of 16 rows of pixels (see Figure B.1). When used to capture full screens, the blocks are read sequentially (starting with block 1 and proceeding to block 12) in 1 millisecond. While one block is being read, the others are being exposed to light. Thus, there is a 1/12 millisecond read time difference between adjacent blocks. Since the different blocks are being read and exposed at slightly different times, image quality can suffer, especially for vertically oriented objects (such as the vibrating vocal folds).

In contrast, the intensified EktaPro allows the exposure time to be much smaller due to the extra gain supplied by the intensifier and the electronic shutter. The blocks are still read in the same manner, but are not exposed to light for nearly the entire frame duration. The electronic shutter on the intensifier allows exposure times to be less than 1/10 that of the unintensified version. This means that all blocks can be exposed nearly simultaneously, resulting in sharper images. As an example, in Figure B.2 an image taken with the normal EktaPro (on the left) is compared to one using the Intensifier VSG. Both are frames from 1000 fps experiments, the highest unintensified frame rate attained in our experiments (there was not enough light amplification without the intensifier to

proceed to higher frame rates). The normal frame has a 1 millisecond exposure time, and the intensified frame a 0.1 millisecond exposure time. It is clear that the intensified frame is sharper and more defined.

The shorter exposure time for the intensified version is therefore important to image quality at these high frame rates. Use of a high gain setting with the shortest exposure time possible given the frame rate desired for a particular experiment will result in clearer images. For many of our experiments the frame rate was 3000 fps (frame period 333 microseconds), with a gain of about 90 (out of 100) and an exposure time of 30 microseconds.

One drawback to using short exposure times is a reduction in the depth of field. Objects will not stay in focus if their distance from the camera varies too much. This can be countered by the use gain in the intensifier, which can add approximately 12 f-stops to the system. However, the gain is limited and a reduction in the depth of field is possible. The amount of reduction depends on the lens used and the distance to the subject. For laryngoscopic applications depth of field did not seem to pose a problem.

1. Increasing the frame rate from 1000 fps

In full-frame mode the maximum frame rate of the EktaPro is 1000 fps. This represents the fastest that the processor can process the information from the camera NMOS imaging sensor. In order to increase the frame rate (or increase the number of frames stored in memory), it is necessary to omit some blocks in the recording process. Thus the frame rate can be increased without increasing the number of bytes the processor has to handle. The number of frames available to record will also increase, yielding equivalent recording times. For example, an EktaPro with enough memory for 1200 full

frames can record for 1.2 seconds at 1000 fps. To record movies at 3000 fps, the number of blocks read must fall to 1/3 of the normal 12. In this case, blocks 5, 6, 7, and 8 are read three times in 1 millisecond to give an exposure time of 333 microseconds. The number of blocks read in 1 millisecond is still 12, so the processor speed does not have to increase. The recording time will still be 1.2 seconds, as the number of recordable frames has tripled along with the frame rate. This technique of using only part of the available resolution may also be used to yield longer recording times at the same frame rate. This results in a tradeoff between field of view, recording time, and frame rate.

For our laryngoscopic experiments 1/3 of the full field (4 blocks out of the 12 available) was enough. Still higher frame rates were possible, but the combination of a small field of view (only 3 blocks for 4000 fps) and low light levels made rates above 3000 fps impractical.

D. Laryngoscopic equipment

Laryngoscopic equipment used to record the images included a 70° rigid endoscope (a Kay Elemetrics, Model 9105), and a flexible nasopharyngoscope (a Pentax FNL-10RP2) used for photographing the larynx via the oral or nasal cavity, respectively. Each scope could be attached to the Kodak camera and to a halogen light source (a Kay Elemetrics Inc. Rhino-laryngeal Stroboscope, model RLS 9100). The light source was typically set on high in order to provide sufficient light for image recording. Even so, as noted, filming at rates above 3000 fps was precluded in part by the limitation in available light. A more powerful light source could have facilitated this situation, but there was some concern that a more intense light source might pose a thermal risk to the subject.

II. Experimental procedure

A. Connecting the equipment together

The experimental setup used for many of our experiments is represented graphically in Figure B.3. For experiments without an intensifier (which we do not recommended) the only difference is the lack of the intensified camera and the recording of "frame mark out" instead of "sync out" onto channel 2 of the A/D.

To record the data, a laptop PC with Labview 4.0 is used in concert with a National Instruments 250 ksps A/D recording at 40 kHz with no prefiltering. The audio is recorded on Channel 0 and is gathered using a powered condenser microphone and a 20 watt amplifier. The amplifier ground is tied in to the oscilloscope chassis ground in order to minimize noise. The GEMS data is input to channel 1 and is also displayed on the oscilloscope to facilitate rapid evaluation of the quality of the electromagnetic signal. Channel 2 is used to record the external sync signal from the intensifier of the EktaPro. Lacking an intensifier, channel 2 can be used to record frame marker information from the EktaPro processor itself. Channels 3 and 4 were used for many things: the subglottal and supraglottal sensors of a Millar esophageal pressure catheter (Model 814), an electroglottograph (EGG), or an inverse-filtered airflow mask.

Synchronizing data and frames

1. Using a TTL trigger to initiate recording

After the subject is in place (seated in a standard ENT exam chair) and the output levels of all input devices have been verified, the data collection process is initiated by pressing a momentary button on a trigger box. This is a simple battery powered circuit

that outputs 0 V DC until the button is pushed, and then outputs 6 V DC. This acts as a TTL high signal for both the Labview-controlled A/D and the EktaPro control unit, indicating that recording should begin. As the EktaPro, once activated, continuously records data in a circular buffer, the TTL start signal notifies it to label the current frame 0 and record until the memory is full (in our case, using 1/3 the screen results in 2456 frames) and then stop. That means that at 3000 fps we can gather about 0.8 seconds of data. At 1000 fps we can gather almost 2.5 seconds of data, which allows us to capture transitions between vocal modes more easily. However, at 3000 fps the pictures are sharper and we gather more information per glottal cycle than at 1000 fps. After the EktaPro records its frames, the result is played back at 30 fps to a television monitor and checked for quality. If it is deemed acceptable, the entire sequence is recorded on a super- VHS VCR. Regular VHS can be used with a penalty in quality, as it has 240 horizontal lines compared to 420 lines for SVHS. The SVHS tape is later digitized for analysis using a MiroVideo DC 30 PCI digitizing card (see Figure B.4 for an example of a digitized frame). Although (as described above) a digital transfer of the EktaPro data is possible using GPIB or serial cable in the form of .tifs, it is a slow process which requires about 40 minutes for one movie of 2456 frames. As we would frequently record 30-40 movies per experiment, this slow speed was unacceptable and the VCR was used in its place. Although the transfer from digital to analog and back to digital will invariably lead to some degradation of image quality (another reason to use SVHS recorders if available), it was simply not feasible to wait 40 minutes to download the data from a single experiment.

2. Recording the external sync or frame marker

In order to determine at what time an exposure occurred, it is necessary to record a frame marker signal from the camera with the same A/D recorder used to sample the audio and other signals. There are two ways to do this - the first is to use the "frame marker" out of the EktaPro control unit, and the second is to use the "external sync" from the intensifier.

The "frame marker" output consists of a single pulse (TTL low to high) emitted at the beginning of each recorded frame. The frame being recorded when the trigger signal is received is defined as frame 0. Its duration is fixed and has nothing to do with the exposure setting of the intensifier. As such, it contains no information on exposure time, which must be calculated from the beginning of each "frame marker" pulse. This is not too difficult, as we will see below, but the long exposure times required by the unintensified EktaPro negate its usefulness with that application. An exposure time of 1 millisecond (the normal exposure time at 1000 fps, the minimum we were able to use due to insufficient light) is about 12% of a 120 Hz glottal period. The image that results from this exposure will be an average of the movement that occurred during this millisecond. The individual images will tend to be blurry, and the only way to clear them up is to use the intensifier or a stronger light source so that the exposure time may be reduced.

The "external sync" pulse from the intensifier, on the other hand, has the advantage of being "high" when the exposure is occurring and "low" when it is not (regardless of the frame rate). Therefore, the place in time at which the exposure occurred can be determined by simply plotting the external sync pulse along with data of interest. The frame being exposed (or read) when the trigger is received is still recorded as frame 0.

All data are recorded at 40 kHz and then everything except the "external sync"

signal is decimated for processing to 10 kHz. The sync signal is left at a higher sampling rate to facilitate the location of the very short (30-100 microseconds) sync pulse in time.

3. Displaying the results

After filtering and decimating the data, the frames are easily correlated with the recorded data by simultaneously plotting the signal of interest with the signal used as the frame marker. If the "ext sync" from the intensifier was used, the process is complete. An example of this is shown in Figure B.5. A section of GEMS data is displayed with the "ext sync" exposure time signal overlaid with the GEMS. Corresponding video frames are shown with arrows referring to their exposure times. The horizontal lines in the video frames are due to a fault in the processor used in the experiment and will not be recorded by a normal processor. This "ext sync" method allows us to rapidly determine to a high degree of accuracy the nature of the GEMS signal at different times in the glottal cycle.

However, if the "frame marker" out of the EktaPro is used, further processing must be done on the "frame marker" signal so that exposure times can be determined. This can be complicated depending on the combination of exposure time and frame rate. The "frame marker" pulse will have to be interpolated to represent the actual exposure time. This is simple enough if the exposure time is the reciprocal of the frame rate, in which case the entire time between frame markers is the exposure time. If, however, the exposure time is shorter than the reciprocal of the frame rate, it is necessary to determine where the exposure occurred.

This can be done if it is understood how the processor captures and stores images. As previously discussed, the NMOS imaging sensor is divided into 12 blocks of pixels (Figure B.1). These blocks are grouped into a "section", the size of which is determined

by the screen size selected. In full screen mode, the section consists of all 12 blocks. At 1/2 screen the section is blocks 4-9, and at 1/3 screen the section consists of blocks 5, 6, 7, and 8. The most blocks the EktaPro processor can read is 12 per millisecond. The amount of screen used thus dictates the minimum exposure time. For example, to capture the full screen using all 12 blocks the maximum frame rate will be 1000 fps. The shortest exposure time is thus 1/1000 of a second. If a shorter exposure time is desired, the number of blocks per frame will have to be decreased.

For example, using only 1/3 of the available screen, the section is made up of four blocks. This section may be read three times per millisecond. Thus the shortest exposure time possible is $1/3 \cdot 1/1000 = 1/3000$ of a second. In this configuration the frame rate may be 3000, 1000, 500, 250, 125, or 50 fps. At 3000 fps each section of four blocks will be recorded and saved as an individual frame. At 1000 fps, only the first section will be recorded and saved - the other two will be discarded. At 500 fps, only every sixth section is recorded and the rest discarded. This process is represented in Figure B.6. Here, an exposure time of 1/3000 is selected and two frame rates are shown. The "frame marker" pulse is graphed vs. time in milliseconds. The numbers indicate which block is being read during the cycle. In this example the trigger signal was received when block 7 was being read, shortly before the frame marker pulse was generated. For the top plot (3000 fps) frame zero was the last section of the previous frame. For the bottom plot (1000 fps) frame zero was the first section read in the previous frame, as sections 2 and 3 of every frame are discarded.

Thus, in order to know when the exposure occurred, you must understand the relation between exposure time, block sections, and frame rate. For each combination of

exposure time and frame rate, the "frame marker" pulse will have to be processed separately. This can be done, but since the exposure times are relatively long with respect to a period, only general correlations of signal and frame images may be made. For those interested in relating the recorded frames precisely to a signal, an intensifier is highly recommended. Without it, most information will have to come from a qualitative analysis of the movies.

4. Limiting factors

The long exposure times required by the unintensified camera are the greatest limiting factor to its use in vocal fold imaging experiments. The images can be blurry and are difficult to relate directly to the other recorded signals.

Another limiting factor is the difficulty of locating the exposure time of each frame due to the discrete nature of digital information and the problems undersampling of continuous signals with digital systems presents us. When recording a 30 microsecond external sync pulse at a sampling frequency of 40 kHz, we are trying to sample a pulse that exists for a time that is comparable to a single sample period. This leads to aliasing and inaccuracy (Porat 1997). For example, at a sampling rate of 40 kHz the sampling period is 25 microseconds. Therefore, any signal with a duration that is not a multiple of 25 microseconds would be represented as a pulse with a duration of a multiple of 25 microseconds (see Figure B.7). This leads to errors in the location of exposure times on the order of 1/2 the sample period. To reduce the magnitude of the errors, a higher sampling rate is needed. Unfortunately, most A/Ds (ours included) are not capable of recording one channel at a different rate than the other channels. Since we are interested in comparing the captured digital frames to another signal such as audio, EGG, or GEMS

data, all of the signals will have to be recorded at the same sampling rate. It was not feasible in our experiments to sample at rates higher than 40 kHz, as we were recording 5 data streams simultaneously and our A/D was only capable of sampling 250 ksamples/sec. Thus, when examining the sampled frame marker signal, we must keep in mind that the actual location of the pulse may be on the order of a 1/2 a sample period away from where it is indicated. Still, at 40 kHz, the maximum error would be about 13 microseconds, and this was acceptable for our experiment.

Conclusions

The Kodak EktaPro camera is capable of capturing sharp images of the vocal folds in operation at 1000 to 3000 fps when using an intensifier. These images are directly comparable to other signals recorded simultaneously and can be located in time to very good accuracy. Without an intensifier, the long exposure times result in individual images that are considerably more blurry, although qualitative evaluations can be done on the resulting movies played back at 30 fps. The ability to study many glottal cycles, one after the other, at many different speeds down to 1 fps can be invaluable, especially when studying transitions in voicing such as register or pitch change.

For details about Kodak's line of high-speed cameras, call Kodak's Motion Analysis Systems Division at 800-462-4307 or go to http://www.masd.kodak.com.

Acknowledgments

The authors gratefully acknowledge Bill Kline at Kodak's Motion Analysis Systems Division for assistance in the theory of operation of the EktaPro and the Intensifier VSG.

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48 and by the University of California at Davis with support from the National Science Foundation.

References

Holzrichter, J.F., G.C. Burnett, L.C. Ng, W.A. Lea (1998). "Speech Articulator Measurements using low power EM-wave sensors" J. Acoust. Soc. Am. **103**, No. 1, 622-625

Porat, B. (1997). *"A Course in Digital Signal Processing"*, pp. 53-57. John Wiley and Sons, New York, NY. ISBN 0-471-14961-6.

Block 12	Rows 177-192
Block 11	Rows 161-176
Block 10	Rows 145-160
Block 9	Rows 129-144
Block 8	Rows 113-128
Block 7	Rows 97-112
Block 6	Rows 81-96
Block 5	Rows 65-80
Block 4	Rows 49-64
Block 3	Rows 33-48
Block 2	Rows 17-32
Block 1	Rows 1-16

239 pixels

Figure B.1. The NMOS imaging sensor is divided into 12 blocks of pixels, each of which can be read separately to increase frame rate.

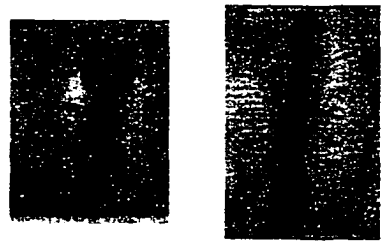
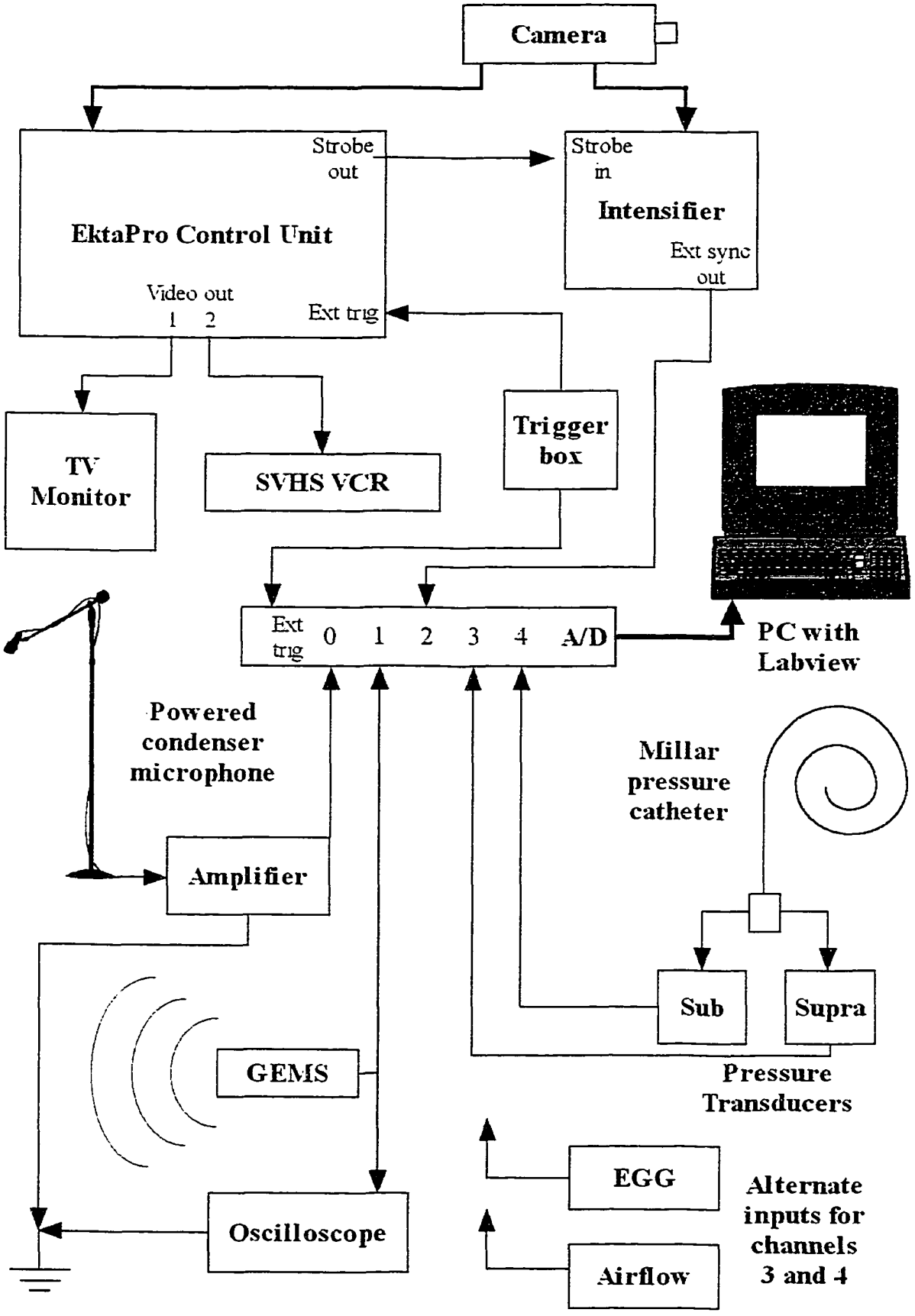


Figure B.2. Comparison between frames taken at 1000 fps. The unintensified (left) frame used a exposure time of 1 millisecond while the intensified (right) frames used an exposure time of 0.1 msec.

Figure B.3. (Next page) Experimental setup for using the intensified EktaPro simultaneously with other data sources.



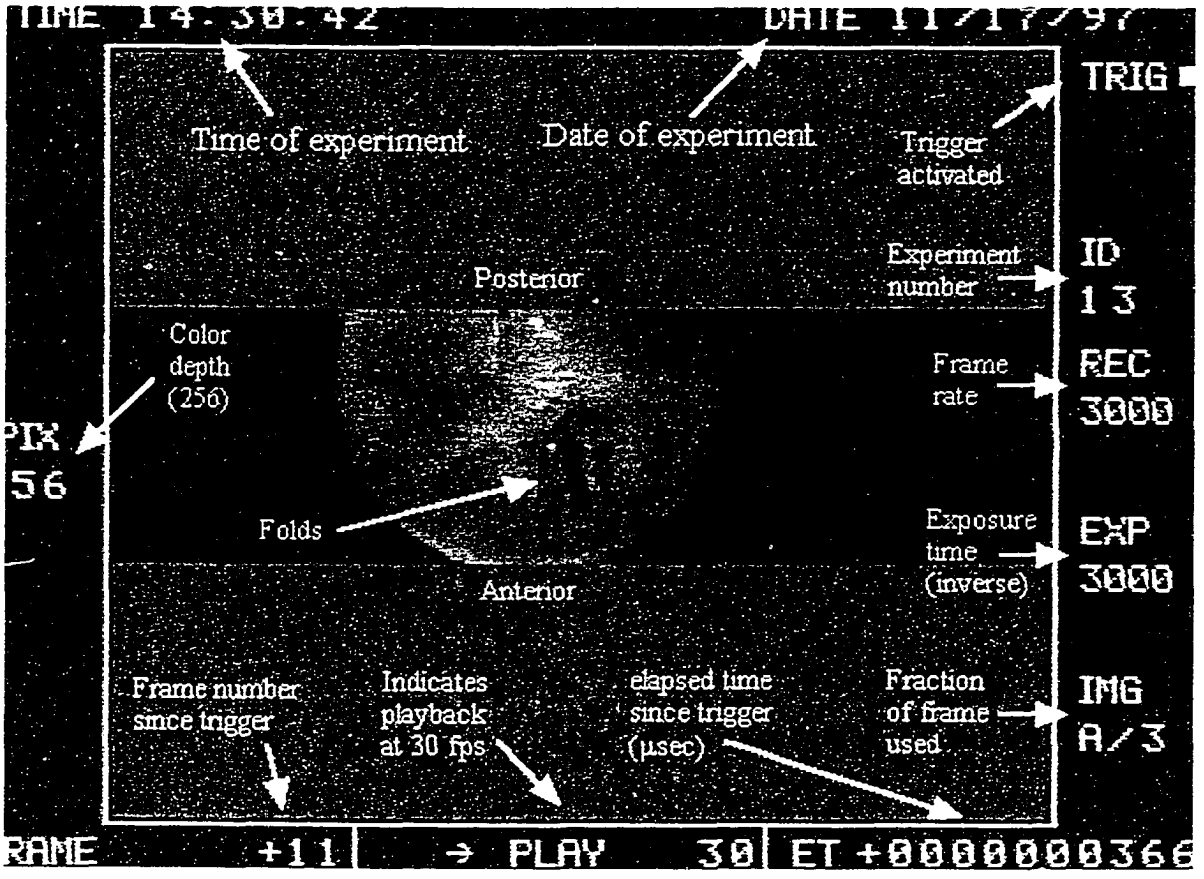


Figure B.4. Digitized frame from an intensified EktaPro using 4 out of the 12 blocks at 3000 fps. The image has been cropped slightly by the digitizing program to reduce file size.

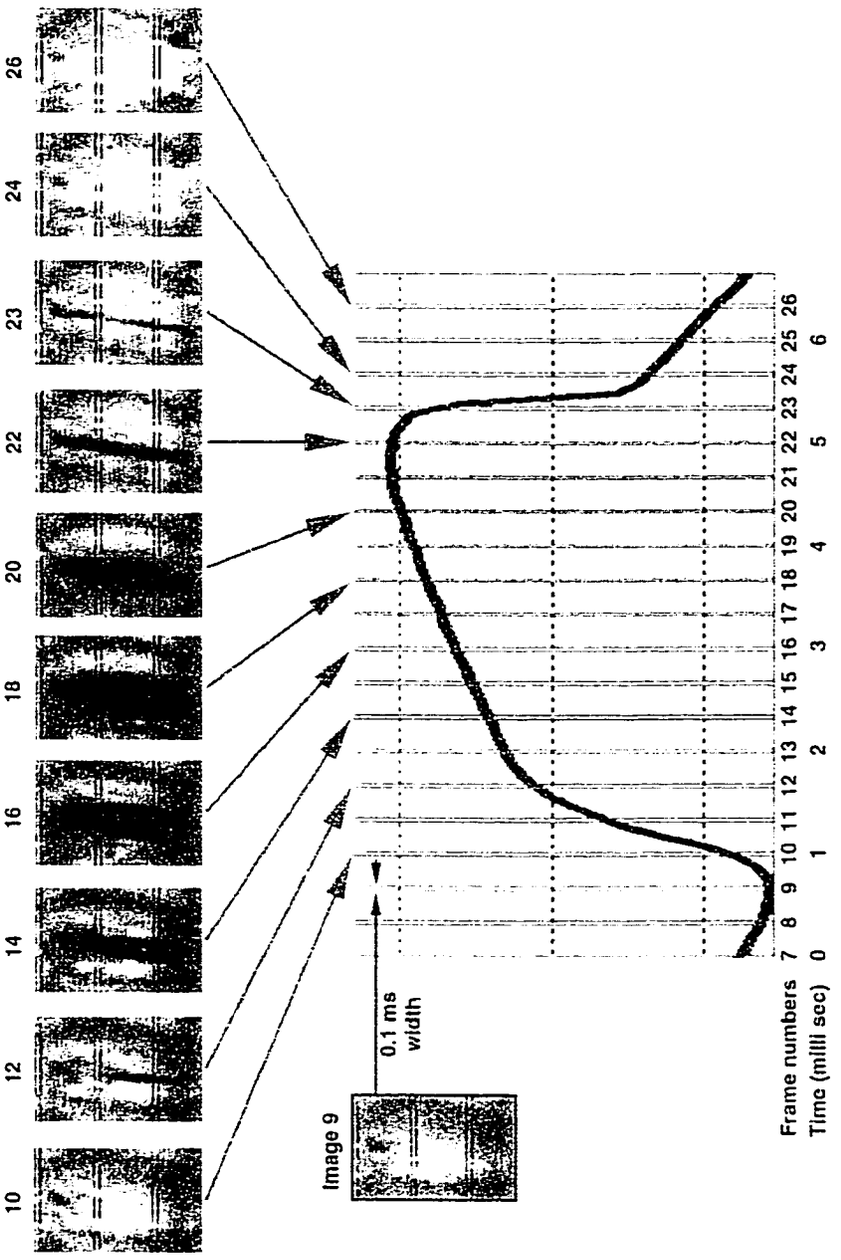


Figure B.5. GEMS signal overlaid with "ext sync" data. The exposure time is much smaller than the frame rate period and this results in sharp images.

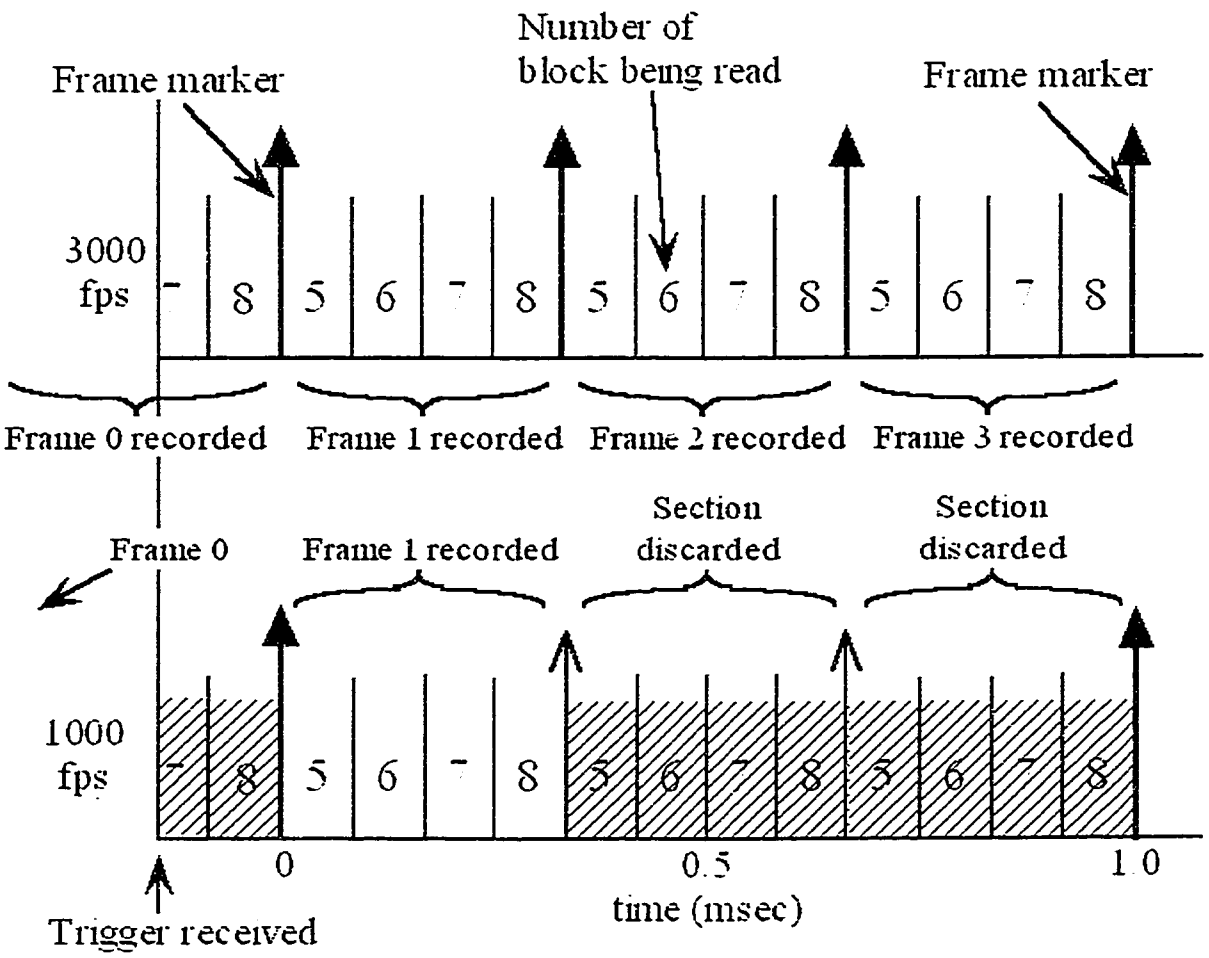


Figure B.6. How to determine where the exposure occurs when the intensifier is not available and the "frame marker" is used. In this example, 1/3 of the screen is used and an exposure time of 1/3000 of a second is selected. The top plot depicts operation at 3000 fps and the bottom at 1000 fps.

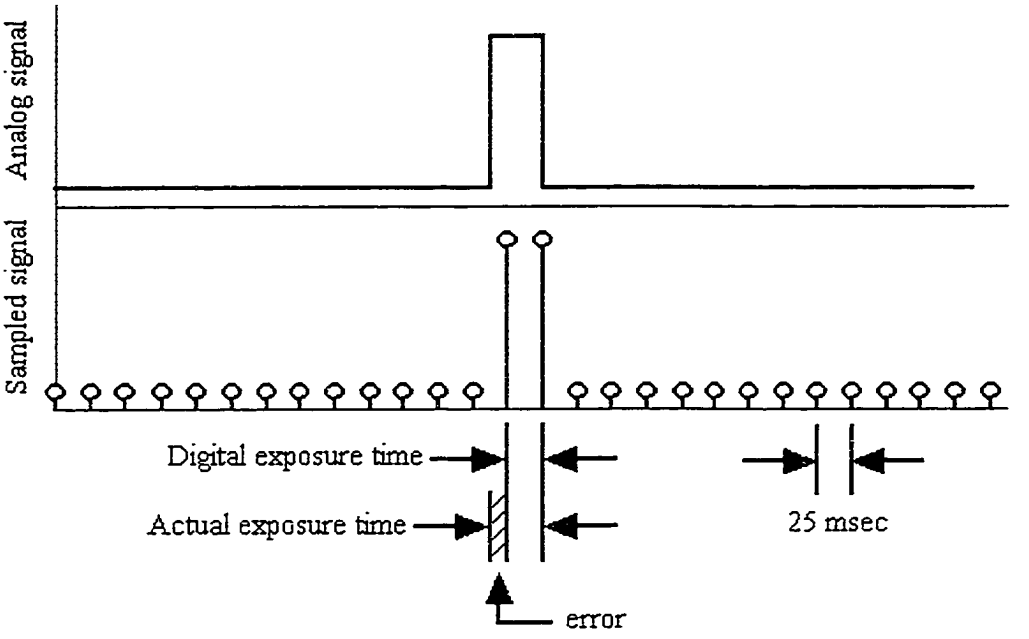


Figure B.7. An example of the error involved when undersampling a fast signal at 40 kHz. The error in locating where the signal occurred depends on the width of the pulse and the sampling rate.

Appendix C.

Accurate and noise-robust pitch extraction using low power electromagnetic sensors

G.C. Burnett, T.J. Gable, L.C. Ng, and J.F. Holzrichter

Lawrence Livermore National Laboratory, POB 808, Livermore CA 94551

Abstract: A new, noninvasive, safe, and robust method of pitch estimation has been developed at the Lawrence Livermore National Laboratory (LLNL) utilizing Glottal Electromagnetic Micropower Sensors (GEMS). They operate in the microwave regime of the EM spectrum at a peak power of less than 1 milliwatt and use a field-disturbance mode of reception in which signals are obtained only from moving tissue. Research has shown that the GEMS signal is strongly correlated with the structure responsible for pitch (the vocal folds), making the signal an excellent source for extremely accurate pitch extraction. The accuracy of the GEMS sensor and corresponding algorithm is validated using tuning forks, synthetic signals, and high-speed video images. It is then compared to two traditional audio-only methods (cepstral and autocorrelation) in normal and noisy environments. This new method using the GEMS and a simple zero-crossing algorithm is shown to be the most accurate, robust, and efficient. These qualities are valuable in many applications such as speaker verification, pitch-synchronous signal processing, noise suppression, pitch training for vocalists, and voice training for the disabled.

EDICS # SA 1.2.2

Corresponding Author: Gregory C. Burnett
Lawrence Livermore National Laboratory
POB 808 L-271
Livermore, CA 94551
Phone: 925-423-3088
Fax: 925-422-7309
Email: burnett5@llnl.gov

I. Introduction

The use of the GEMS in relation to speech applications was first explored by Holzrichter *et al.* [1]. The authors demonstrated how the GEMS and other EM sensors could be used to measure vocal articulator motion in real time for speech characterization. One particularly interesting speech characteristic (measurable by the GEMS) is the closing of the vocal folds during their oscillation. This is normally defined to be the beginning of the voiced cycle and also known as a "glottal close", as the glottis is defined as the space between the folds. The time between one glottal close and the next is referred to as the pitch period, and the reciprocal of the period is the pitch.

A. The Sensor

The GEMS are derived from a general class of Micropower Impulse Radars (MIR), invented by Tom McEwan at LLNL in 1993. Information on the particulars of the operation of this class of sensor may be found in [2] and [3]. The specific sensor used in these experiments was modified to filter out low frequency motion (with a 3-dB frequency of about 70 Hz) and to operate in the near field so as to be sensitive to vocal fold motion-induced vibrations. Other MIR sensors have been modified to detect lower frequencies of motion (up to a few Hz) and these have been used for observation of the jaw and tongue [1]. The GEMS transmits about 10 cycles of a 2.3 GHz wave at a pulse repetition frequency of 2 MHz. It waits a specified (adjustable) time and then mixes the return with a delayed version of the transmitted signal (homodyne detection). The time it waits (the range gate) and the antenna radiation pattern define a "bubble" of sensitivity within which motion is detected. The pulse trains are on the order of a few nanoseconds (ns) long and are repeated every 500 ns, so multiple reflections or overlapping returns from different trains do not cause problems. The GEMS uses a filtering and averaging method so that only changes in the reflectivity of objects in its bubble of sensitivity generate a signal. If there is no motion, there is no AC signal. In essence, the GEMS signal closely mirrors the physical motion of a moving

surface if the extent of motion is much less than the wavelength of the pulses (about 12 cm in air for this sensor) and is of sufficiently high frequency (i.e. above ~70 Hz).

B. Use of the sensor in this study

In these experiments, the GEMS antennae (simple rectangular copper foils about 1.5 cm by 0.8 cm) were positioned just below the laryngeal prominence behind a thin (2 mm) plastic case in light contact with the skin (see Figure C.1). The GEMS may be moved farther away from the skin (up to about 6 cm for the present sensor; work is underway to increase that further) and may also be moved up and down the trachea, but to facilitate comparison between individuals the above placement was used.

An example of the GEMS signal along with the corresponding audio is shown in Figure C.2. Notice the rapid fall in the GEMS signal - this is when the vocal folds are in the process of closing. This has been verified through the use of high-speed (3000 frames per second) digital video. A Kodak EktaPro EM 1012 with an Intensified Imager VSG was used with a normal laryngoscope to obtain several 2456-frame videos of the vocal folds in motion. Audio and GEMS signal were recorded simultaneously with the exposure-time signal from the EktaPro controller. At 3000 fps, three frames are recorded in a millisecond. Each frame was exposed for only 30 microseconds, resulting in many clear "snapshots" of the folds for each glottal cycle. Figure C.3 shows the GEMS signal for a single glottal cycle and some of the corresponding video frames. Since the rapid fall occurs as the folds are closing, where it crosses zero (the signal is AC coupled) is defined as the beginning of the glottal cycle. The audio signal feature corresponding to the closure of the folds is detected by a conventional condenser microphone about 1.4 msec later. This delay is due to the travel time of sound through the vocal tract and to the microphone (about 50 cm).

Pitch calculation algorithms

There are many pitch calculation algorithms in use today [4], [5], [6], but all rely on the audio signal alone to determine pitch. The GEMS provide us with a simpler way to find pitch through a direct measurement of the motion of physical structures surrounding the vocal folds. The three methods we chose to compare for this paper are a simple zero crossing (for the GEMS signal), autocorrelation (time domain), and the cepstral (frequency domain).

A. Zero crossings using the GEMS signal

The zero crossing algorithm is quite simple – the zero crossings are calculated by determining where the normalized GEMS signal changes from positive to negative. There is only a single such crossing per cycle regardless of the phoneme being voiced as long as the register does not lapse into vocal fry (for an explanation of vocal registers see [7]). The positive to negative crossing is chosen as it defines the beginning of the glottal cycle (see Figure C.3) and is quite linear as it passes through zero, facilitating the use of linear interpolation for increased accuracy. To obtain the pitch, the length of the signal (in samples) from one zero crossing to the next is determined and then translated to time by division by the sampling frequency. In most experiments two glottal cycles are averaged so that the "pitch windows" are two cycles long, but any number of cycles may be used. The period calculated for the two cycles is inverted to get pitch in frequency:

$$\text{pitch} = \frac{2}{\text{period}} = \frac{2 \cdot \text{sampling frequency}\left(^{\text{samples}}/_{\text{sec}}\right)}{\text{length of signal (samples)}}$$

The GEMS pitch algorithm (a block diagram is given in Figure C.4) uses a fixed window of 35 ms to begin processing the signal. Within that 35 ms window, it looks for enough of a signal to indicate that voicing has occurred. It does this by dividing the summed absolute amplitudes of each normalized window by the length of the window in samples to get the average amplitude. If the average amplitude of the window is below a threshold, a zero pitch is assigned (denoting silence or unvoiced speech) and the window is moved to the next 35 ms of data. If, however, the average amplitude exceeds the threshold, the window is considered voiced. The GEMS is very

stable and quiet due to the bandpass filter between 70 Hz and 7 kHz, so voiced speech detection errors (both false positives and negatives) are quite rare. Unless the speaker is voicing (and thus the folds and surrounding tissue vibrating), the GEMS produce very little signal. As the turn-on time for the folds is usually quite rapid, voicing onset and offset may be reliably detected to within a few milliseconds. This ability to measure near-instant voicing onset and offset in a compact, low-power package would be quite useful in speech recognition and coding.

If the window is considered voiced, then a subroutine finds the position in time of all the zero crossings in the window. The first zero crossing is defined as the beginning of the "voiced speech" window, and the pitch calculated from that first zero crossing to the third one. Linear interpolation of the zero crossings is used to increase accuracy, as near the zero crossing the GEMS signal is quite linear. The third zero crossing then becomes the beginning of the next window. Thus the windows are fixed in length at 35 msec until voicing is detected, and then the pitch of the speech dictates the window length. This allows us to do pitch-synchronous processing, enabling more accurate Fourier transforms and decreasing the number of calculations required. If the window is determined to be unvoiced, no further processing is needed.

B. Autocorrelation

We used the clipped autocorrelation pitch detection algorithm described by Rabiner [8]. The data is first lowpass filtered with a 99-point linear phase FIR filter. It is then segmented into 30 ms rectangular windows which are stepped 10 ms at a time, resulting in a 20 ms overlap. Each window is tested to decide if it is voiced or unvoiced by an energy calculation, which is compared to a threshold. If it passes the threshold then the window is center clipped to 68% of the maximum. The autocorrelation is then computed and the location of the first peak that is 30% or more of the correlation at zero is considered to be the pitch period.

C. Cepstral

The cepstral method uses the Fourier transform (FT), but it is not a purely frequency based algorithm as it uses the FT twice to get back into a type of time space [4]. The real (as opposed to complex) cepstral proceeds as follows:

$$X = \Im(x)$$

$$X_2 = \log_{10}(X \cdot X^*)$$

$$\text{Cepstrum} = \Im(X_2) \cdot \Im(X_2)^*$$

The cepstral independent variable exists in the time domain and is known as the "quefrency". In essence, the magnitude of the second power spectrum will have peaks at quefrencies that correspond to repeated peaks in the first power spectrum. The first peak location of sufficient amplitude is defined to be the pitch. Thus a signal with a fundamental at 100 Hz and harmonics every 100 Hz will have peaks in X_2 every 100 Hz, and will have a single peak in the cepstrum at a quefrency of 10 ms. As the cepstral involves finding the power spectrum of a power spectrum, it needs a good number of harmonics in order to be effective. In our experiments, we used 40 ms Hamming windows with a 10 ms step, thus a 30ms overlap.

Both conventional methods use peak finding to determine the pitch period. Finding the discrete peak is rather inaccurate at a sampling rate of 10 kHz. By taking the difference of the cepstral or autocorrelation vector as a first approximation to the derivative, we can interpolate and determine where the derivative would cross zero if it were linear. Using this zero crossing we can approximate where the peak is in-between the sampled points.

For both acoustic methods it is also necessary to smooth the initial pitch contours as there can be large deviations in the calculated pitch. For this we used Rabiner's standard smoothing algorithm [8] utilizing a 3,5 median filter and then a simple Hann linear filter. The GEMS pitch contour (due to its inherent stability and natural acoustic noise immunity) requires no smoothing.

III. Accuracy, stability, and robustness of the GEMS signal and algorithms

A. Tuning fork measurement

To demonstrate the accuracy of the GEMS signal, the motion of a vibrating tuning fork was measured with the GEMS and compared to the simultaneously recorded audio. A PC laptop with Labview 4.0 and an A/D from National Instruments were used with a 10 kHz sampling frequency. All analysis was done with Matlab 5.1. Both the GEMS and the microphone were about 1 cm away from a vibrating tine. A portion of the normalized data is shown in Figure C.5. The audio signal is offset by +1 in the y direction to facilitate comparison. The GEMS mirrors the audio signal of the tine almost perfectly. We calculated the pitch using the GEMS zero-crossing algorithm every 2 cycles (about 6.6 ms) for 3 seconds on both the GEMS signal and the audio signal (this is permissible because the audio signal was a pure sinusoid). For the audio the mean was 329.2 Hz with a standard deviation of only 0.1 Hz. The GEMS signal also yielded an average of 329.2 Hz, with a standard deviation of 0.3 Hz. Thus, the experiment shows the GEMS is capable of excellent accuracy even at the low sampling rates commonly used for speech (10 kHz). At higher sampling rates the accuracy would increase, as we are locating an event in time rather than frequency space, and with higher sampling rates comes better time resolution.

B. Comparison of the GEMS algorithm to the autocorrelation and cepstral methods

1. Synthetic signal

In order to directly compare the GEMS method to the audio-only methods outlined above, a synthetic signal s(t) was constructed with a variable fundamental frequency and 15 harmonics of the fundamental (as the cepstral algorithm must have a significant number of harmonics to be effective):

$$s(t) = \sum_{n=1}^{15} \frac{1}{N} \sin\left(2\pi N f_k t\right)$$

The fundamental frequency f_k was varied from 80 Hz to 300 Hz and the sampling rate was defined to be 10 kHz. The length of s(t) was 100 msec and each method was allowed to find as

many pitch points as it could in that time period. The mean of those points was defined as the pitch at that frequency, $p(f_k)$. The relative error was defined as

$$e(f_k) = \frac{|f_k - p(f_k)|}{f_k}$$

The percent error vs. frequency plot is shown in Figure C.6. Note how the GEMS algorithm has the lowest level of error across almost the entire spectrum. The results are similar if interpolation is not used. Interpolation increases the accuracy by about the same amount for all methods.

In addition to being more accurate, the GEMS method also has a significant advantage in terms of cost of computation. In the above example (which was calculated on Matlab 5.1), the number of CPU flops (floating point operations) required to calculate the pitch was computed for all methods. The results are illustrated in table 1, along with the mean of the standard deviation and the error across all f_k for the synthetic signal. It is clear that the GEMS method requires far less computational power and is both more accurate and precise than either of the conventional methods. For signals with unvoiced portions, the GEMS cost would be even lower as the GEMS algorithm does no processing on unvoiced windows. The other two methods must process the entire data stream, as they have no information about when the voiced speech begins or ends.

Method	Cepstral	Autocor	GEMS
# kflops (average)	1100	1800	8
% error (average)	0.27	0.043	0.0083
Std. dev. (average)	0.071	0.153	0.052

Table C.1. Number of kflops required to determine the pitch for a 100 ms synthetic signal in Matlab 5.1, the average error in pitch, and the average standard deviation from the synthetic pitch (80 to 300 Hz) for the three methods.

This low computational cost is due to the simplicity of the time domain-based zero crossing algorithm, which is possible because the signal from the GEMS is clean, relatively simple (with a sharp feature), and is unaffected by acoustic noise.

Speech in quiet and noisy environments

In this noise sensitivity experiment, two male subjects were recorded speaking both a single vowel (/i/) and a sentence ("When all else fails, use force"). The audio and GEMS recording were performed at 40 kHz with no prefiltering on the same equipment used for the tuning fork experiment. The data was then filtered and decimated (using a distortion free digital filter) to 10 kHz. One subject (the "speaker") was approximately 30 cm from the microphone and the GEMS was lightly touching the centerline of the neck directly below the laryngeal prominence. The second subject (the "noise") was seated approximately 60 cm from the microphone, approximating a 12 dB signal to noise ratio. The second subject spoke along with the first but delayed his onset in order to illustrate the difficulties experienced by acoustic-only pitch algorithms when a second speaker is present. All three methods were used to determine the pitch. The results are shown in Figures C.7 and C.8. Note how in the absence of noise (at the beginning of the speech), the three pitch contours are quite similar; in contrast, the introduction of the second (noisy) speaker is quite noticeable in the acoustic methods, which exhibit large errors. The GEMS contour is unaffected. Absolute acoustic noise rejection is one of the foremost attributes of the GEMS method.

IV Conclusion

The GEMS approach has been shown to provide superior pitch information at a very low computational cost compared to that obtained through conventional (audio-only) means. The comparisons were made under ideal conditions – with a stressed speaker or rapid vibrato (where the pitch can change significantly over a few glottal cycles), the instantaneous pitch information available with the GEMS algorithm would be even more advantageous. It is immune to acoustic

noise and is non-invasive, safe, and portable. A "double-boom" configuration in which a headset with a microphone boom is modified with a second boom containing the GEMS antenna is being constructed and is one possible method of commercial implementation.

Potential applications for the GEMS' unique characteristics include vocal training, training for the deaf, vocal stress detection, and prosodical supplementation for speech recognition engines (especially useful for tonal languages). The latter has been advocated in the past [9], but has never been implemented. The GEMS zero crossing algorithm makes low cost, very accurate and robust pitch information (as well as the location of voicing onset/offset times) available for many applications.

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48 and by the University of California at Davis with support from the National Science Foundation.

References

[1] Holzrichter, J.F., Burnett, G.C., Ng, L.C., and Lea, W.A. (**1998**). "Speech articulator measurements using low power EM-wave sensors", J. Acoust. Soc. Am. **103**(1), 622-625.

[2] McEwan, T.E. (**1994**). U.S. Patent No. 5,345,471 (1994), U.S. Patent No. 5,361,070 (1994).

[3] McEwan, T.E. (**1996**). U.S. Patent No. 5,573,012 (1996)

[4] Noll, A. (**1966**). "Cepstrum pitch detection," J. Acoust. Soc. Am. 41, 293-309.

[5] Rabiner, R., Cheng, M., Rosenberg, A. and McGonegal, C. (**1976**). "A comparative study of several pitch detection algorithms" IEEE trans. on acoustics, speech and signal processing Vol. 24, 399-418.

[6] Rabiner, R. and Juang, B. (**1993**). "*Fundamentals of speech recognition*" (Prentice Hall, New Jersey)

[7] Titze, Ingo R. (**1994**). "*Principles of Voice Production*" (Prentice-Hall, Englewood Cliffs, NJ)

[8] Rabiner, R., Sambur, M., and Schmidt, C. (**1975**). "Applications of nonlinear smoothing algorithm to speech processing" IEEE trans. on acoustics, speech and signal processing, Vol. 23, 552-557.

[9] Lea, Wayne A. (ed.) (**1989**). "*Toward Robustness in Speech Recognition*" (Speech Science Publications, Apple Valley MN), see pp. 117-118, 499.

Figures

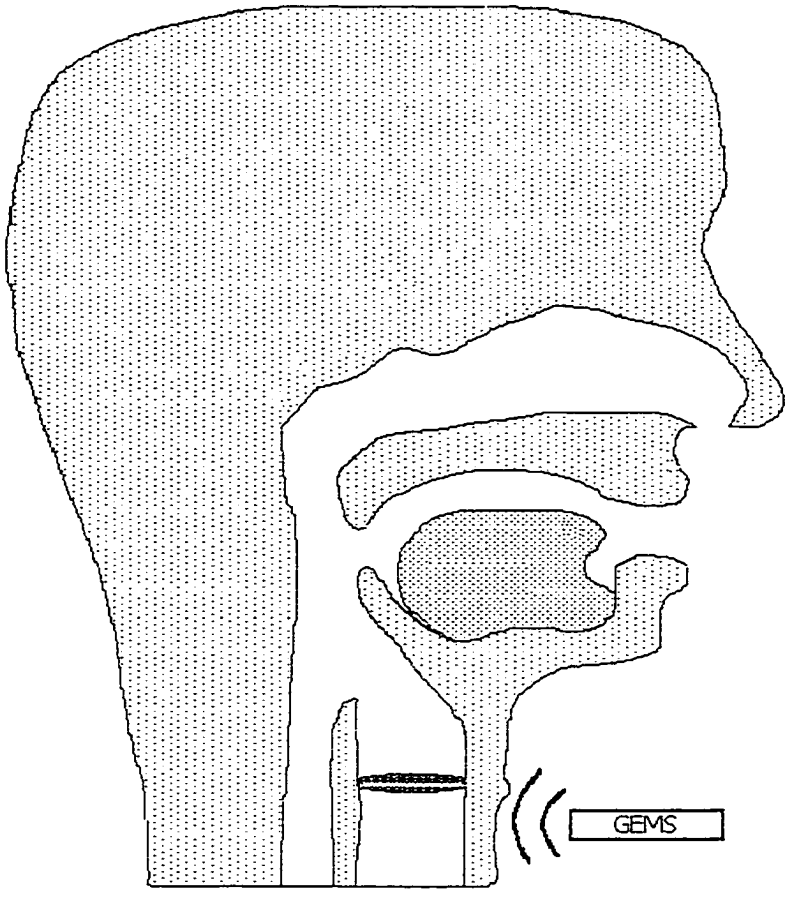


Figure C.1. GEMS placement for pitch measurements.
Normally light skin contact is made but is not necessary.

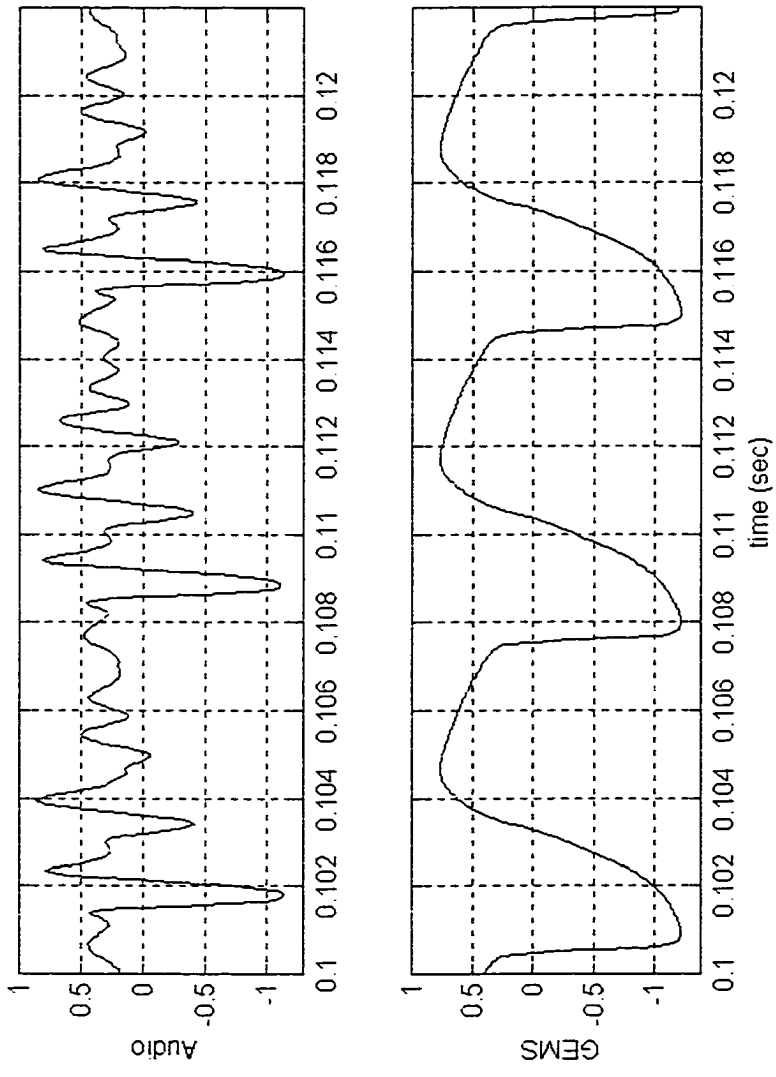


Figure C.2. Audio and GEMS signals from 29 year old male native English speaker, voicing */a/* ("ah")

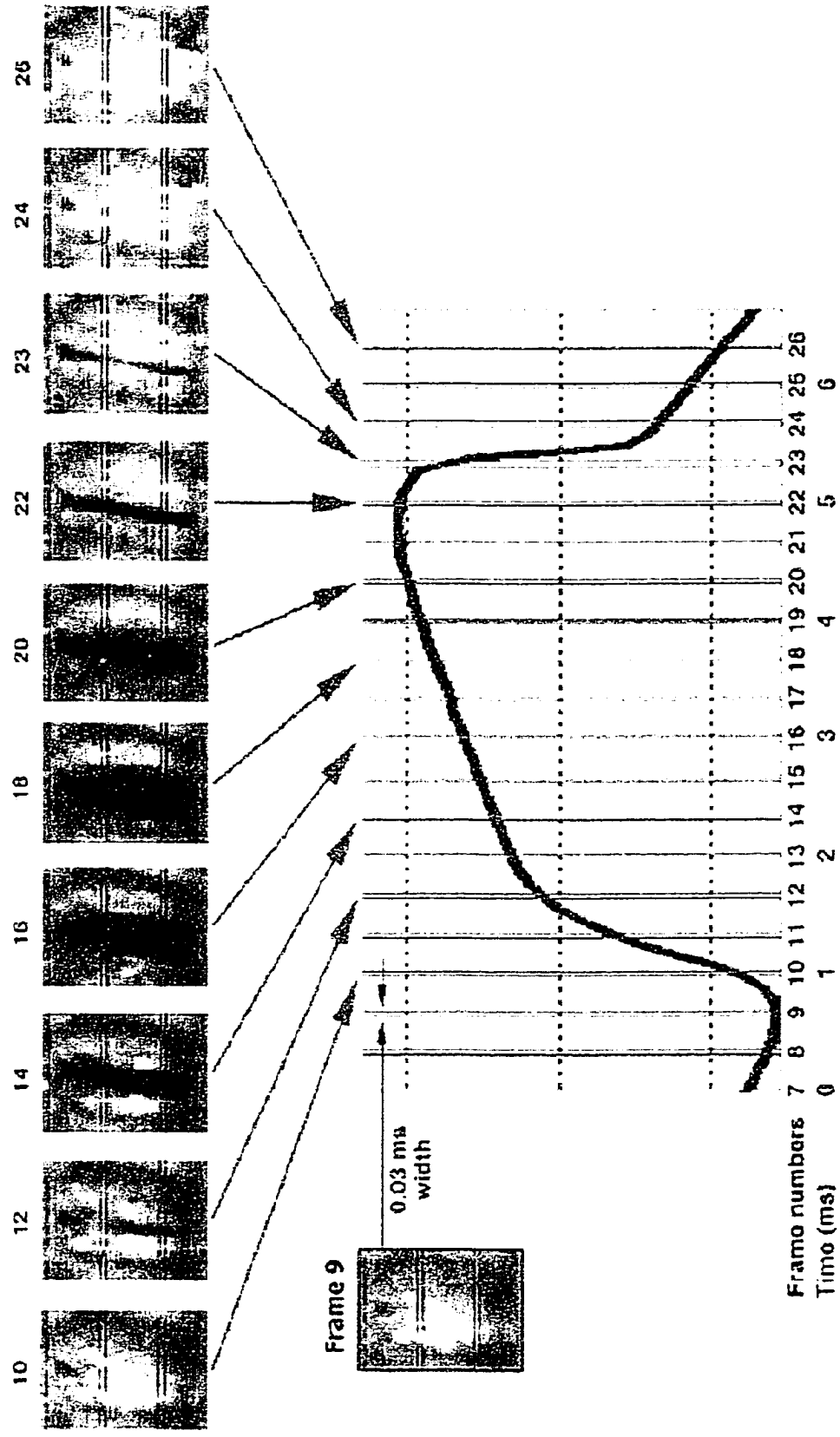


Figure C.3. GEMS signal overlaid with the corresponding high-speed vocal fold video frames. Each bar is 30 microseconds wide and represents the exposure time of the frame.

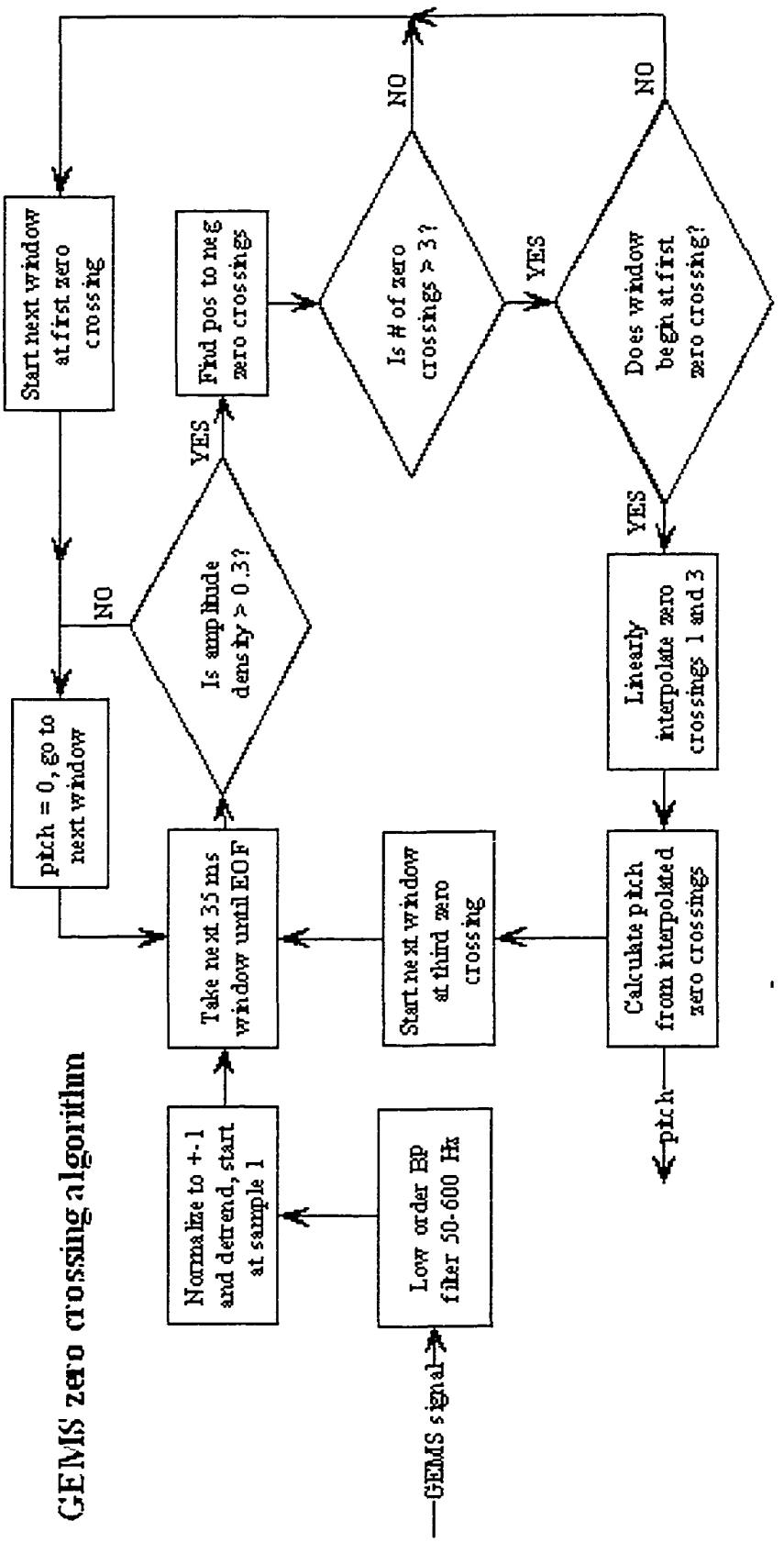


Figure C.4. Block diagram of the GEMS zero-crossing algorithm.

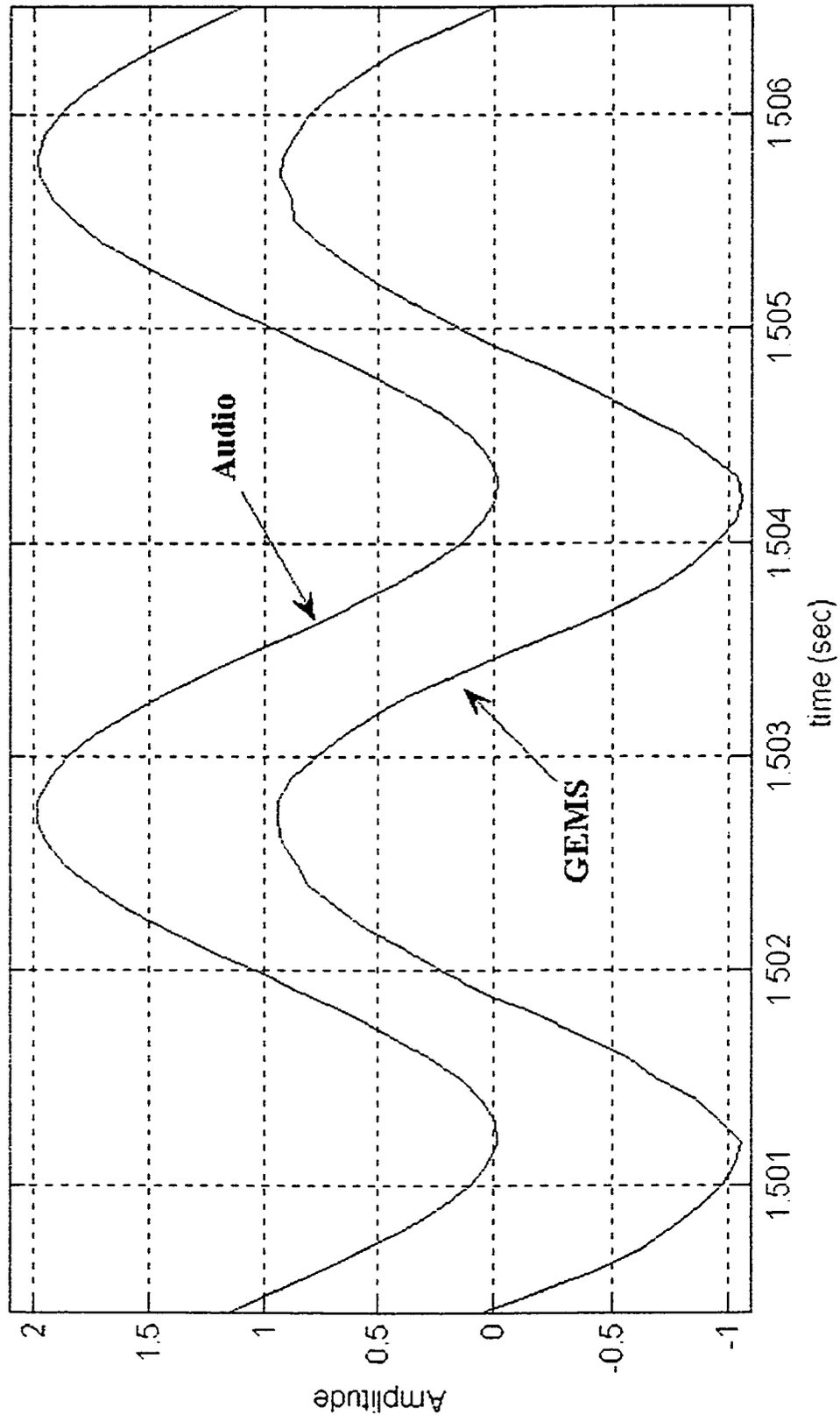


Figure C.5. Normalized signals from a tuning fork. The audio (upper) is offset in the y direction to facilitate comparison to the GEMS signal (lower).

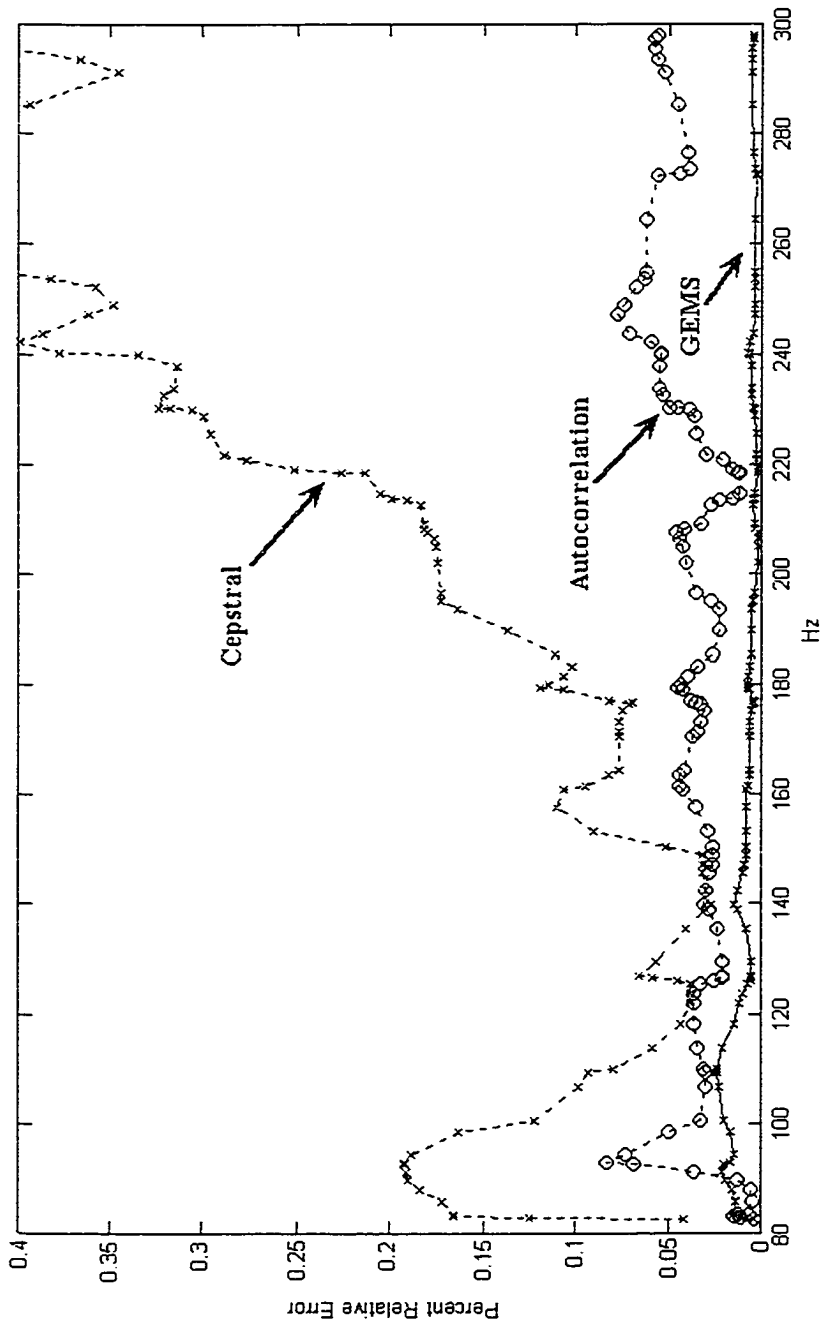


Figure C.6. Relative error vs. actual pitch for each pitch algorithm. A three second long synthetic signal with multiple harmonics was used. Cepstral (-x), autocorrelation (-o), and GEMS (x).

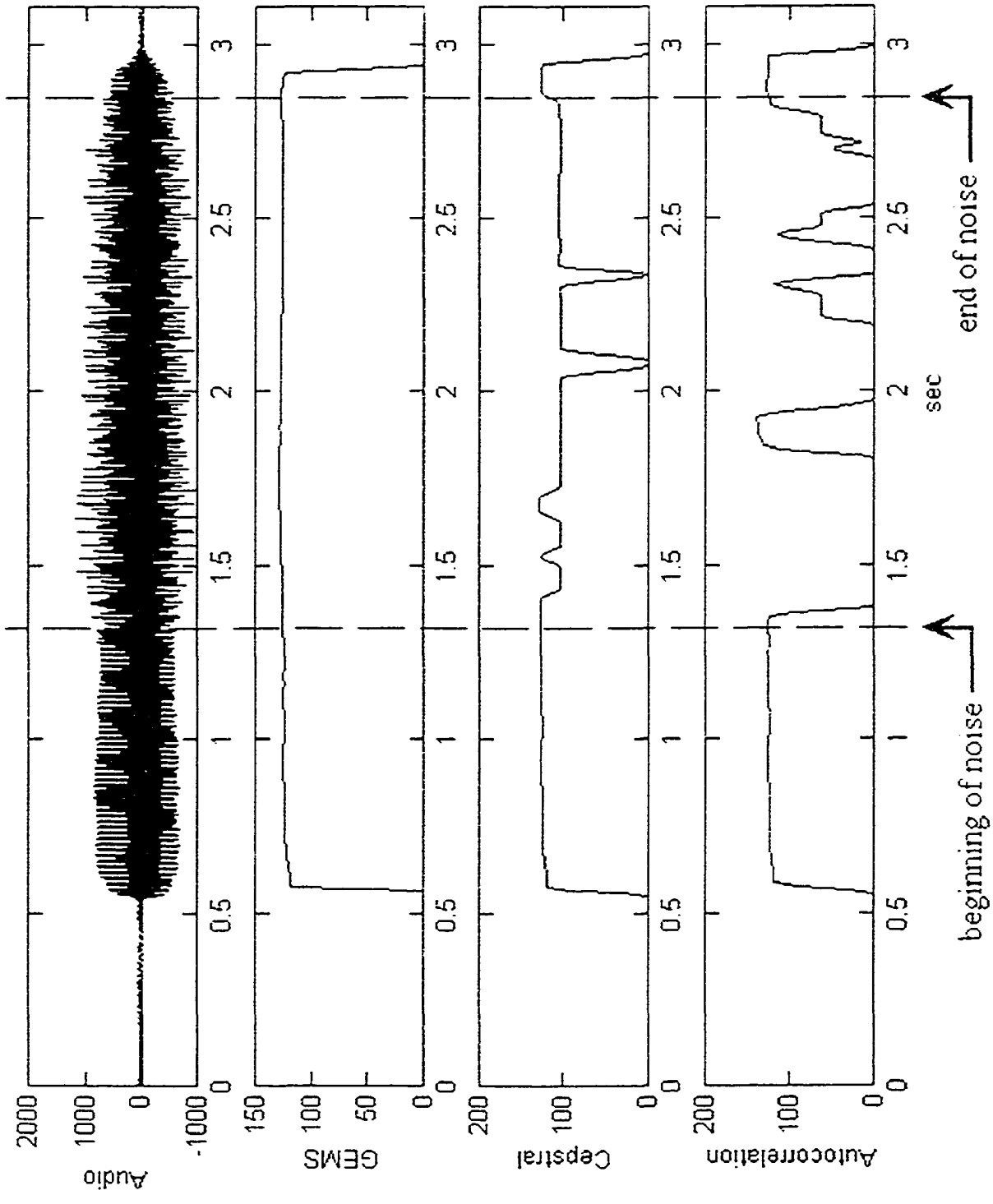


Figure C.7. Noisy (includes a second male speaker) audio signal (/i/) with pitch contours for GEMS, cepstral, and autocorrelation methods. The GEMS signal is unaffected by the noise.

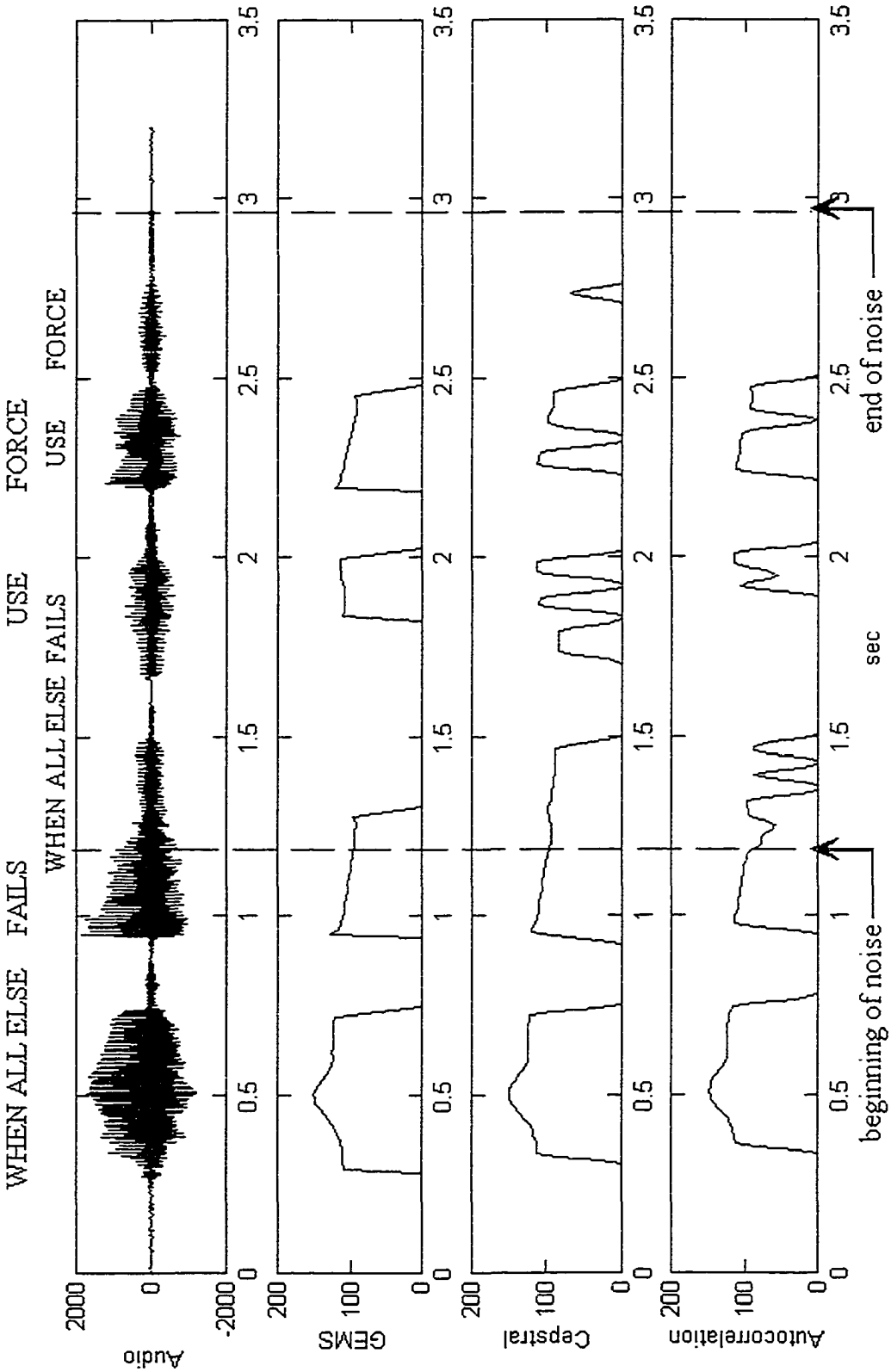


Figure C.8. Noisy (includes a second male speaker) audio signal ("When all else fails, use force") with pitch contours for GEMS, cepstral, and autocorrelation methods. Again the GEMS signal is unaffected.

Appendix D. The Phonetic Alphabet for American English (from Titze 1994)

Phoneme #	Vowel sounds	Key words
1.	i	each, free, keep
2.	I	it, bin
3.	e	ate, made, they
4.	ɛ	end, then, there
5.	ae	act, man
6.	a	ask, half, past
7.	ɑ	alms, father
8.	upside-down a	hot, odd, dog, cross
9.	reversed c	awl, torn
10.	o	obey, note, go
11.	ʊ	good, foot
12.	u	ooze, too
13.	upside-down e	alone, among, circus, system
14.	upside-down e	father, singer
15.	ʌ	up, come
16.	reversed ɛ	urn, third

Phoneme #	Consonant sounds	Key words
17.	p	pie, ape
18.	b	be, web
19.	m	me, am
20.	w	we, woe
21.	upside-down w	why, when
22.	f	free, if
23.	v	vine, have
24.	θ	thin, faith
25.	δ	then, clothe
26.	t	ten, it
27.	d	den, had
28.	n	no, one
29.	l	live, frill
30.	r	red, arrow
31.	s	see, yes
32.	z	zoo,
33.	ʃ	show, ash
34.	3	measure, azure
35.	j	you, yes
36.	UC	huge, human
37.	k	key, ache
38.	g	go, big
39.	UC	sing, long

40.	h	he, how

Phoneme #	Vowel combos (dipthongs)	Key words
41.	eI	aid, may
42.	aI	aisle, sigh
43.	reversed-cI	oil, joy
44.	aʊ	owl, cow
45.	oʊ	own, go

Phoneme #	Consonant combos (affrictates)	Key words
46.	tʃ	chew, each
47.	d3	gem, hedge

UC = unreproducible character, unknown font. See Titze (1994) for symbol.

Appendix E. Glossary

3-dB frequency: Frequency at which the magnitude response of a filter is three decibels below the response in the passband. Used as a design and performance metric for filters.

AC coupled: Electrical term that denotes the connection of a circuit through a capacitance so that only changing (AC) signals are passed. DC signals are rejected.

Allpass filter: Filter with $|H(z)| = 1$ of the form:

$$H(z) = \frac{a^* + z^{-1}}{1 + az^{-1}}$$

Can be used to stabilize unstable filters while retaining the desired magnitude response.

Anterior (anatomical description): In front of.

Antennae: Plural form of antenna.

Anticausal filter: a filter whose impulse response h[n] = 0 for n > 0. Has negative phase. Cannot be used in real time; all processing must be done offline.

Antisymmetric: For an FIR filter, it is antisymmetric if $h[n] = -h[n_1 + n_0 - n]$ where n_1 is the position of the last filter coefficient and n_0 is the position of the first. For causal filters this reduces to h[n] = -h[N - n] where N + 1 is the total number of taps.

Autodyne detection: Radar detection where the radar wave is emitted and the reflected or backscattered signal is detected. No phase information is available. This method has a lower signal to noise ratio than coherent (homo- or heterodyne) detection. The name autodyne is used by comparison to the other two techniques. The technique is also called "direct detection," or "incoherent detection."

Bernoulli effect: The effect on a fluid due to velocity or pressure change. Based on the conservation of energy. Potential (pressure) and kinetic (velocity) energy add to get the total energy, which is conserved. An increase (decrease) in velocity results in a decrease (increase) in pressure. Discovered by Swiss scientist Daniel Bernoulli.

Bistatic: Antenna(e) configuration in which the transmit and receive antennae are either the same physical unit or are located very near one another. The electromagnetic energy is transmitted from the transmitter, reflects from the medium of interest, and travels back to the receiver. See monostatic.

Cascading filters: Arranging filters one after the other so that a combined filter of higher order is constructed. The response of the new filter is then

$$H(z) = H_0(z) \cdot H_1(z) \cdot \ldots H_N(z)$$

where H(z) is the new filter constructed using N cascades.

Causal filter: a filter whose impulse response h[n] = 0 for n < 0. Can be used in real time to filter an incoming signal. If FIR, has linear phase.

Dynamic Time Warping: Used to time-align speech patterns to account for differences in speaking rates across speakers and for repetitions of patterns for the same speaker. Depends heavily upon knowledge of speech onset and offset. (Rabiner & Juang, p. 201-239)

EGG (Electroglottograph): Instrument that conducts radio frequency (MHz) AC current from one electrode to another (bistatic configuration). An electrode is placed on either side of the trachea by the laryngeal prominence. The current is passed through the folds when they are in contact and is greatly restricted when

they are not. The resulting signal is proportional to the amount of vocal fold contact area.

Field disturbance mode: The mode in which the GEMS operates. Any disturbance in the reflectivity of the objects in the GEMS FOV results in a signal from the GEMS.

FIR filters: Finite impulse response filters. Can be built using a finite number of taps and delays. Have linear phase and can be any causality. Four main types classified on: even or odd number of taps N (not counting zero, total number of taps is N + 1), and whether it is symmetric or antisymmetric around its central point. It goes in order from even and symmetric to odd and antisymmetric.

Type i	Type 2	Type 3	Type 4
Even and symmetric	Odd and symmetric	Even and antisymm.	Odd and antisymm.

Formant: Speech community vernacular for resonance. Used most often in the description of the resonance locations of the vocal tract.

FOV: Field of view.

Frequency response: The Fourier transform of the impulse response of a filter.

$$H(e^{jw}) = \sum_{n} h[n] \cdot e^{-nj\omega}$$

Frontal plane (anatomical description): A plane that bisects the body into front and rear halves, running from left to right. Also known as coronal.

Glottis: The space between the vocal folds. Defines the glottal area as the area surrounding the vocal folds.

Heterodyne detection: The heterodyne configuration is similar to the homodyne, except that the mixed signal is derived from a separate source, normally offset in frequency from the transmitted signal.

Hidden Markov Models: A statistical algorithm used in modern speech recognition. The statistical method differs from the template by virtue of the way the speech is perceived: Instead of a well-defined pattern existing for every word, phoneme, or sentence, the speech signal is characterized as a parametric random process whose parameters can be estimated in a precise, well-defined manner. The theoretical basis for the operation of these recognizers is decades old and exceedingly complex (Rabiner & Juang, p. 321-389).

Homodyne detection: In the homodyne method, the reflected signal is mixed with the transmitted signal. This results in a signal from the receiver which is proportional to the phase difference between the two signals.

Impulse response: The output from a filter when the input is the delta function, a single impulse at n (or t) = 0. It is equal to the coefficients of the filter h[n]. The transform of the impulse response is the frequency response if in ω space ($H(e^{jw})$) or the transfer function if in z space (H(z)).

Inferior (anatomical description): Below

Laryngeal prominence: The bump on the upper anterior thyroid cartilage, commonly know as the "Adam's apple".

Linear phase: A characteristic of FIR filters. The phase is linearly related to ω and causes no distortion of the signal.

Linear Time Invariant (LTI): The most common type of filter or modeling process used due to its simplicity and robustness. Linear refers to the reaction of the plant to the input – it is linearly related. Mathematically, it is linear if for $u[n] = x[n] + y[n]$, $U(z)H(z) = X(z)H(z) + Y(z)H(z)$. It is time invariant if a shift in the input results in a corresponding shift in the output: $x[n-k] \rightarrow H(z) \rightarrow y[n-k]$.

Magnitude envelope: Term used in this thesis to describe the magnitude of the GEMS return vs. the phase difference in the transmitted and received signals.

Magnitude response: The magnitude of the response of a filter vs. frequency. For example, an allpass filter has a magnitude response equal to 1 at all frequencies, while a lowpass filter will have a response near 1 below its cutoff frequency and near 0 above it.

Minimum Phase: Used to describe a filter that has all its poles and zeros inside the unit circle. A minimum phase filter has a stable inverse.

Monostatic: Antennae configuration in which the transmit and receive antennae are separated by a significant distance. The electromagnetic energy is transmitted from the transmitter through the medium of interest and to the receiver. See bistatic.

Noncausal filter: a filter consisting of causal ($h[n] = 0$ for $n < 0$) and anticausal ($h[n] = 0$ for $n < 0$) sections cascaded together. Cannot be used in real time, all processing must be done offline. Can have zero phase if symmetric.

Phase response: The phase shift vs. frequency that occurs when a signal is acted on by a filter. Each frequency component of the input will be shifted by some amount, which can result in a distortion of the input. Even allpass filters have a nonzero

phase response and will distort an input without changing the magnitude of its frequency components.

Phoneme: A building block of language. Each language can be subjectively divided into a number of linguistically distinct speech sounds. Each one is a phoneme. Examples in American English are /i/ (beat) and /ɑ/ (father).

Plant (signal processing): The system (or combination of many systems) that modifies an input signal and yields an output signal.

Posterior (anatomical description): Behind.

Sagittal plane (anatomical description): A plane that bisects the body into right and left halves and runs from back to front.

Sensitivity envelope: The variation of the magnitude of the GEMS signal vs. distance to a reflector, assuming the index of refraction of the reflector is greater than that of the incident medium ($n_2 > n_1$).

Skin depth: A measure of the penetration capability of EM waves. As EM waves penetrate any material, energy is lost due to the conductivity of the material. The distance at which the wave has lost 1/e (about 37%) of its energy is termed the skin depth.

Standard deviation: Statistical method of explaining how far a member of a population is from a mean of that group. At least 75% of the values will fall within two standard deviations of the mean and at least 89% will fall within three standard deviations of the mean. Mathematically

$$\sigma_x = \sqrt{\frac{(x-\bar{x})^2}{N}}$$

where $\bar{x}$ is the mean of the population with N members.

Subglottal: The region of the trachea directly below the glottis. Defined as inferior to the laryngeal prominence (Adam's Apple).

Superior (anatomical description): Above.

Symmetric: For an FIR filter, it is symmetric if $h[n] = h[n_1 + n_0 - n]$ where n_1 is the position of the last filter coefficient and n_0 is the position of the first. For causal filters this reduces to $h[n] = h[N - n]$ where $N + 1$ is the total number of taps.

Template Model: Speech recognition model in which speech is divided into typical sequences of speech frames for a pattern (such as a word) via some averaging procedure, and to rely on the use of local spectra distance measures to compare patterns. Each pattern can be referred to as a speech vector and the vectors compared to find the best fit. Patterns used include phonemes, diphones, words, and sentences. Dynamic time warping is used to time-align patterns to account for differences in speaking rates across speakers and for repetitions of patterns for the same speaker.

Transfer function: The z transform of the impulse response of a filter, where $z = r \cdot e^{jw}$.

$$H(z) = \sum_n h[n] \cdot z^{-n}$$

In essence, it describes the behavior of the filter at all relevant digital frequencies.

Transverse plane (anatomical description): A plane that splits the body into upper and lower halves, running from front to back.

Unvoiced speech: Speech that is formed without the use of the vocal folds.

Variance: The square of the standard deviation.

Vocal Folds: The flaps of multilayered tissue located in the trachea directly behind the laryngeal prominence, or "Adam's Apple" as assigned by myth. Used to produce

sound by modulating airflow from the lungs. Commonly known as the "vocal cords".

Voiced Speech: Speech that is formed using the vocal folds.

Appendix F. References

Alipour-Haghighi, F., and Titze, I.R. (1991). "Elastic models of vocal fold tissues" J. Acoust. Soc. Am. **90(3)**, 1326-1331

Baer, T., Gore, J.C., Gracco, L.C., and Nye, P.W. (1991). "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels" J. Acoust. Soc. Am. **90**, 799-828

Boerner, W.M. and Chan, C. (1986) "Inverse methods in electromagnetic imaging", *Medical applications of microwave imaging*. New York: IEEE Press, ISBN: 0-87942-196-7, pp.213-227.

Bogert, B.P. and Peterson, G.E. (1957). "The acoustics of speech". In L.E. Travis (ed) *Handbook of Speech Pathology* (pp 109-173). Appleton-Century-Crofts, New York

Burdette, E.C., Cain, F.L. and Seals J. (1986). "In-Situ Tissue Permittivity at microwave frequencies: Perspective, Techniques, Results". *Medical applications of microwave imaging*. New York: IEEE Press, ISBN: 0-87942-196-7, pp.13-40.

Chan, K.H., and Lin, J.C. (1987). "Microprocessor-based cardiopulmonary rate monitor". *Medical and Biological Engineering and Computing*, Vol. 25, pp. 41-44.

Dang, J., Honda, K., and Suzuki, H. (1994). "Morphological and acoustical analysis of the nasal and the paranasal cavities" J. Acoust. Soc. Am. **96(4)**, 2088-2100

Den Boom, Van and Den Enden, Van. "The Determination of the Orders of Processes and Noise Dynamics" 1974

Duck, F.A. (1990). "*Physical Properties of Tissue*" (Academic Press Inc., San Diego, CA)

Eriksson, L.J. (1980). "Higher order mode effects in circular ducts and expansion chambers" J. Acoust. Soc. Am. **68**, No. 2, 545-550.

Faber, Robert Y. Jr., and Rybinski, Valerie A. "UTP Cabling and the Effects of EMI" White Paper. The Siemon Company, Watertown, Connecticut.
http://www.siemon.com/references/tech/presentation.html

Fant, G. (1960). "*The Acoustic Theory of Speech Production*" (Mouton, The Hauge)

Flanagan, J.L. (1965). "*Speech Analysis Synthesis and Perception*" (Springer-Verlag, NY).

Flanagan, J.L., Ishizaka, K., and Shipley, K.L. (1975). "Synthesis of speech from a dynamic model of the vocal cords and tract". The Bell System Technical Journal, Vol. 54, No. 3, March 1975.

Flanagan, J.L. and Landgraf, L. (1968). "Self oscillating source for vocal tract synthesizers". *IEEE Trans. on Audio and Electroacoustics*, AU-16, 57-64.

Gauffin, J., Binh, N., Ananthapadmanabha, T.V. and Fant, G. (1983). "Glottal geometry and volume velocity waveform" In D. Bless and J. Abbs (Eds.), *Vocal Fold Physiology: Contemporary research and clinical issues* (pp. 194-201). College-Hill Press, San Diego CA.

Griffiths, David J. (1989). "*Introduction to Electrodynamics*" (Prentice-Hall, Englewood Cliffs, NJ) 2nd edition

Haddad, W.S., Rosenbury, E.T., Johnson, K.B., and Pearce, F.J. (1997). Measurements of the Dielectric Properties of Body Tissues and Fluids at Microwave Frequencies (Lawrence Livermore National Laboratory, to be published)

Halliday, D., Resnick, R., and Krane, K. (1992). "*Physics*" (4th ed, vol. 1). (John Wiley & Sons, Inc. New York, NY)

Hartmann, William J. (1997). "*Signals, Sound and Sensation*" (American Institute of Physics, Woodbury, NY)

Holzrichter, J.F., G.C. Burnett, L.C. Ng, W.A. Lea (1998). "Speech Articulator Measurements using low power EM-wave sensors" J. Acoust. Soc. Am. **103**, No. 1, 622-625

Ishizaka, K. and Matsudaira, M. (1972). "Fluid mechanical considerations of vocal cord vibration". SCRL Monograph 8.

Ishizaka, K., French, J.C., and Flanagan, J.L. (1975). "Direct determination of vocal tract wall impedance". IEEE Trans. On Acoustics, Speech, and Signal Processing, Vol. ASSP-23, No. 4, August 1975.

Jacobi, J.H. and Larsen, E.L. (1986) "Linear FM pulse compression radar techniques applied to biological imaging", *Medical applications of microwave imaging*. New York: IEEE Press, ISBN: 0-87942-196-7, pp.138-147.

Kent, R.D. (1978). "Imitation of synthesized vowel by preschool children" J. Acoust. Soc. Am. **63**, 1193-1198.

Kent, R.D. (1979). "Isovowel lines for the evaluation of vowel formant structure in speech disorders" J. of Speech and Hearing Disorders, 44, 513-521.

Kinsler, L. and Frey, A. (1962). "*Fundamentals of Acoustics*" (2nd ed). (Wiley, New York)

Kinsler, L., Frey, A., Coppens, A., and Sanders, J. (1982). "*Fundamentals of Acoustics*" (3rd ed). (John Wiley and Sons, New York)

Land, D.V. (1995). "Medical microwave radiometry and its clinical applications". *IEE Colloquium on "The Application of Microwaves in Medicine"*, Digest No. 1995/041, London, UK.

Larsen, E.L. and Jacobi, J.H. (1986). "Methods of active microwave imagery for dosimetric applications", *Medical applications of microwave imaging*. New York: IEEE Press, ISBN: 0-87942-196-7, pp.118-137.

Lea, Wayne A. Editor (1986). "*Trends in Speech Recognition*" (Speech Science Publications, Apple Valley, MN)

Lin, JC (1986). "Microwave propagation in biological dielectrics with application to cardiopulmonary interrogation", *Medical applications of microwave imaging*. New York: IEEE Press, ISBN: 0-87942-196-7, pp.47-58.

Meaney, P.M, Paulsen, K.D., Hartov, A., and Crane, R.K. (1996) "Microwave imaging for tissue assessment: Initial evaluation in multitarget tissue-equivalent phantoms". *IEEE Trans. on Biomed. Engin.*, Vol. 43, No. 9, pp. 878-890.

McGowan, R (1991). Phonation from a continuum mechanics point of view. In J. Griffin and B. Hammarberg (Eds), *Vocal Fold Physiology: Acoustic, perceptual, and physiological aspects of voice mechanisms* (pp. 65-72). Singular Publishing Group, Inc., San Diego, CA.

Milenkovik, P., and Mo, F. (1988). "Effect of the vocal tract yielding sidewall on inverse filter analysis of the glottal waveform" " J. Voice, 2(4), 271-278.

Mizushina, S., Ohba, H., Abe, K., Mizoshiri, S., and Sugiura, T. (1995). "Recent Trends in Medical Microwave Radiometry". *IEICE Trans. Commun.*, Vol E78-B, No. 6, pp. 789-798.

Moll, K.L. (1965). "Photographic and radiographic procedures in speech research" ASHA Reports, No. 1 Proceedings of the Conference on Communication Problems in Cleft Palate, 129-140

Moore, C.A. (1992). "The correspondence of vocal tract resonance with volumes obtained from magnetic resonance images" JSHR, 25,1009-1023

Morse, P.M., and Ingard, K.U. (1968). "*Theoretical Acoustics*", McGraw-Hill.

Murphy, Q.M. (1994). "Radar Tomography: A new concept in medical imaging". *Annals of Dentistry*, Summer 1994, 53(1), pp. 5-14.

Noll, A.M. (1966). "Cepstrum pitch determination". *J. Acoust. Soc. Am.* V41(2), 293-309.

Narayanan, SN; Alwan, AA and Haker, K. (1995). "An articulatory study of fricative consonants using magnetic resonance imaging". *J. Acoust. Soc. Am.* V98 (3), 1325-1347.

Peterson, G.E. and Barney, H.L. (1952). "Control methods used in a study of vowels" J. Acoust. Soc. Am. **24**, 175-184

Porat, B. (1997). *"A Course in Digital Signal Processing"* John Wiley and Sons, New York, NY. ISBN 0-471-14961-6.

Rabiner, Lawrence, and Juang, Biing-Hwang (1993). *"Fundamentals of Speech Recognition"* (Prentice-Hall, Englewood Cliffs, NJ)

Rschevkin, S.N. (1963). *"A Course of Lectures on the Theory of Sound"* (Pergamon Press, New York, NY. LOC #62-19271.

Scherer, R., and Titze, I.R. (1983). Pressure-flow relationships in a model of the laryngeal airway with a diverging glottis. In D. Bless and J. Abbs (Eds.), *Vocal Fold Physiology: Contemporary research and clinical issues* (pp. 179-193). College-Hill Press, San Diego CA.

Skolnik, M.I. (1986). "Radar measurements, resolution, and imaging of potential interest for the dosimetric imaging of biological targets", *Medical applications of microwave imaging*. New York: IEEE Press, ISBN: 0-87942-196-7, pp.59-65.

Story, Brad H. (1995). "Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract" (PhD. Thesis, University of Iowa).

Story, BH; Titze, IR and Hoffman, EA. (1996). "Vocal tract area functions from magnetic resonance imaging". *J. Acoust. Soc. Am.* V100 (1), 537-554.

Story, BH and Titze, IR. "Parameterization of vocal tract area functions by empirical orthogonal modes". *NCVS Status and Progress Report*, November 1996 9-23.

Sulter, A.M., Miller, D.G., Wolf, R.F., Schutte, H.K., Wit, H.P., and Mooyaart, E.L. (1992). "On the relation between the dimensions and resonance characteristics of the vocal tract: A study with MRI" Mag. Res. Imag., 10, 365-373

Svirsky, M.A., Stevens, K.N., Matthies, M.L., Manzella, J., Perkel, J.S., and Wilhelms-Tricario, R. (1997). "Tongue surface displacement during bilabial stops", J. Acoust. Soc. Am. **102(1)**, 562-571.

Titze, Ingo R. (1983). "Mechanisms of sustained oscillations of the vocal folds" In I. Titze & R. Scherer (Eds.) Vocal Physiology: Voice Production, mechanisms and functions (pp.349-357). (Denver Center for the Performing Arts, Denver, CO)

Titze, Ingo R. (1984). "Parameterization of the glottal area, glottal flow, and vocal fold contact area" J. Acoust. Soc. Am. **75(2)**, 570-580.

Titze, Ingo R. (1988). "The physics of small amplitude oscillation of the vocal folds" J. Acoust. Soc. Am. **83(4)**, 1536-1552.

Titze, Ingo R. (1994). "*Principles of Voice Production*" (Prentice-Hall, Englewood Cliffs, NJ)

Titze, I.R., Story, B.H., Burnett, G.C., Holzrichter, J.H., Ng, L.C., Lea, W.A. (1999). "Comparison between electroglottography and electromagnetic glottography", under review by JASA.

Young, J.D. and Peters, L. (1986). "Examination of video pulse radar systems as potential biological exploratory tools", *Medical applications of microwave imaging*. New York: IEEE Press, ISBN: 0-87942-196-7, pp.82-105.