REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
Dec. 2004

3. REPORT TYPE AND DATES COVERED
Final 5/4/04-12/4/04

4. TITLE AND SUBTITLE
Voiced Excitations

5. FUNDING NUMBERS
HR0011-04-C-0073
P00002

6. AUTHOR(S)
Holzrichter, J.F.: Ng, L.C.; Steinkraus, R.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
SNH Instruments and Systems, LLC
200 Hillcrest Rd.
Berkeley, CA , 94705

8. PERFORMING ORGANIZATION REPORT NUMBER
NA

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
DARPA / ATO Attn: Dr. Lisa Porter
3701 North Fairfax Drive
Arlington, VA 22203-1714

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
NA

11. SUPPLEMENTARY NOTES
NA

12a. DISTRIBUTION / AVAILABILITY STATEMENT
UNCLASSIFIED UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

To more easily obtain a voiced excitation function for speech characterization, measurements of skin motion, tracheal tube, and vocal fold motions were made and compared to EM sensor-glottal derived excitation functions during voiced speech. Skin motions were measured at 9 locations on the neck and face. Typically 2 micron amplitudes were observed depending upon the sound and location. The location below the vocal folds provided the most robust information. Tracheal wall motions above and below the glottis were measured with a low power EM interferometer attached to a narrow beam antenna. Motions of 5-10 microns were measured with best results by using the E field polarized in the vertical direction at the subglottal location. Tracheal motion and vocal fold signals are quite different in amplitude, in spectral response, and in E-field polarization response, from each other. Both skin and trachea signals showed spectral detail up to about 600 and 800 Hz respectively. Methods to use these data to generate sufficiently good excitation functions of voiced speech are discussed.

14. SUBJECT TERMS
Radar & EM Speech,
Voiced Speech Excitations

15. NUMBER OF PAGES
61

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
UNCLASSIFIED

18. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED

20. LIMITATION OF ABSTRACT
UL

"Voiced Excitations" FINAL REPORT
Darpa Contract 1- HR0011-04-C-0073 P0002
SNH Instruments and Systems, LLC
May 4, 2004 to Dec. 4, 2004

Investigators:
John F. Holzricher, Lawrence C. Ng, Robert Steinkraus
SNH Instruments and Systems, LLC

This Final Report includes the following topics:

Task Objectives for determining voiced excitation functions for purposes of vocoding in high noise environments:

Task 1: Estimation of voiced excitation functions using skin surface vibration measurements.
Task 2: Estimation of voiced excitation functions using tracheal wall vibration measurements using EM sensors.
Task 3: Estimation of voiced excitation functions using EM sensor measured vocal fold opening versus time.
Task 4: Modeling and summary of voiced excitations determined using data from Tasks 1-3.

The report covers:
Technical Problems
General Methodology
Technical Results
Important Findings and Conclusions
Significant Hardware Development
Special Comments
Implications for Further Research

Sponsored by:
Defense Advanced Research Projects Agency
Advanced Technology Office/ATO
Program: VOCODER
ARPA Order No. S455/00 Program Code: 3410
Issued by DARPA/COM under Contract No. HR0011-04-C-0073 P00002

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the Defense Advanced Research Projects Agency or the U.S. Government

Table of Contents:

“Voiced Excitations”
Darpa Contract 1- HR0011-04-C-0073 P0002
SNH Instruments and Systems, LLC
May 4, 2004 to Dec. 4, 2004

Abstract:

Air pressure or air flow excitation functions of voiced speech have been difficult to estimate using acoustic techniques, especially in real time in noisy situations. Except for pitch estimation, these data are not usually used for fielded electronic speech characterization such as in automatic speech recognizers, vocoders (e.g., speech compression), speech synthesizers, and related applications. Recently, by using small electromagnetic-wave radar-like sensors (i.e., EM interferometric sensors), it has become possible to measure the movements of several types of tissues, both those comprising the vocal folds and those comprising the walls of the vocal tract (cheaply and in real time). Tissues such as the vocal folds, the vocal tract walls, and the overlying skin surface all move in response to glottal induced air pressure pulsations during voiced speech. It was the purpose of this contract to determine to what degree these motions can be used to work backwards, using inversion techniques, to determine a sufficiently good voiced excitation function for Vocoding in High Noise Environments.

The work on “Voiced Excitations” shows that the surface vibrations of neck skin, during voicing, carry a great deal of information describing the excitation function of voiced speech. It is shown that EM and optical interferometric sensors, which measure to 10 kHz, obtain skin vibration excitation limited to below 600-800 Hz. The spectral sensitivity is observed in the PSD of the skin vibration and also in the skin transfer-function high frequency roll-off, which relates internal air pressure pulsations to surface vibration. While there is additional acoustic information, present interferometric or doppler sensors are limited in sensitivity by their measuring mode (detecting velocity or position vs time rather than acceleration), and by the actual sensor “noise floors”.

Finally, skin surface vibrations measurements can be corrupted by external loud noise, depending upon the intensity and frequency content of the noise. Loud noise (>110dB) at a typical frequency of 200 Hz, can cause 0.1 to 0.5 micrometer skin motion, which is about 10% of the measured skin motion due to internal voiced pressure pulses, or in other words, the noise is at –10dB. Loud, spectrally complex sounds cause a skin motion response spectrum proportional to the pressure/unit frequency of the noise.

Interior vocal tract wall tissue also moves in response to internal pressure pulsations, but is unaffected by external noise. Its motions can be measured by directional EM sensors, which provide spectral information on the excitation that is limited to below 600 Hz for super-glottal wall motions, but extends up to about one kHz for sub-glottal wall motions. Additional information on vocal fold excitations regarding best EM sensor coupling, best antennas—dipole, patch, waveguide, used in a monostatic or bistatic mode, and other properties are described.

A methodology for determining a voiced excitation function, using limited information from skin surface, internal tract wall surfaces, and vocal fold motions, as well as training data, is described. It makes use of the established methods of first obtaining glottal spectral data using EM interferometric sensors (see procedures

described at http://speech.llnl.gov/), then converting these data to an air flow or an air pressure excitation function, then performing noise removal from the corresponding acoustic speech signal, then finding the vocal tract formants, and finally performing subsequent tasks for the speech application at hand. This methodology makes use of the fact that except for pitch and amplitude, a user’s excitation function shape is relatively constant over a period of months of time (using the same EM sensor/antenna system). If a user’s excitation function is measured using even a simple EM sensor (e.g., GEMs mode) during a training session, then the excitation shape data can be registered and subsequent real-time data, with lower spectral fidelity, can be used to continually update the excitation function during field use. A second approach is to use the somewhat limited, but very informative spectral information from tissue motions, to select a “best fit” excitation from a catalog of standard (prior measured or calculated) pitch glottal functions, which are then warped to fit the pitch period, fall times, and widths of the user’s shape.

The voiced excitation functions derived from the procedures and data developed during this report enable the definition of very good excitation functions. The acquisition of these functions can be obtained according to the needs of the speech applications. In the case of Darpa’s Vocoding in High Noise Programs, excellent denoising and speech compacting can be accomplished using the procedures described in this report.

I. Introduction to “Voiced Excitations”

In the mid 1990s, a group at the Lawrence Livermore National Laboratory (Holzrichter, Burnett & Ng 1998) showed that low power electromagnetic sensors, similar to radars, could be used to measure what appeared to be vocal fold motions during voiced speech (Holzrichter et al 1999). Subsequent work showed that indeed the motions of vocal folds and several other tissues in the vocal tract could be measured, quantified, and used to enhance automatic speech characterizations (see data at the web site http://speech.llnl.gov). Recently Darpa began a program to demonstrate a low bit rate Vocoder (i.e., a speech compressor) to be used in high noise environments exceeding 110 dBc (see Darpa BAA04-35). This and other vocoding programs are enabled if EM sensors or other speech recording instruments can provide sufficiently accurate, uncorrupted, excitation function information to meet the data-quality requirements needed by the noise removal and speech compression algorithms. It has been shown that if sufficiently broad voiced spectral information is obtained, see Ng et al (2000) for denoising and Ng et al (2002) for vocoding, the Darpa Vocoding in High Noise program can be accomplished. However, at present obtaining this data is more difficult than is desired, because an EM sensor must be placed within about 1 cm of the vocal-folds’ location (just below the Adam’s apple, noted as position 5 in this report). For field use, it may be better to obtain data from another location on the neck or face.

A pressure function of voiced speech is formed as the vocal folds close rapidly, abruptly interrupting air flowing from the lungs through the vocal folds, into the vocal tract, and out through the lips (and also nose and skin surfaces). This interruption of airflow generates a “negative” pressure step function in the air flow (ΔP is about - 5 cm H_2O), causing a negative pressure step with wide spectral content to propagate up the vocal tract. The articulator-defined cavities in the vocal tract form an acoustic filter, which filter the pressure spectrum (and also air-flow spectrum), defining the frequencies of the voiced speech sound unit. This pressure change also causes sound waves to propagate upward through the mouth, as well as downward toward the lungs, and also outward through the vocal tract wall tissue to the skin, where it is radiated. In addition, quasi-static expansion and contraction of the vocal tract wall tissues occur in response to low frequency air pressure changes (a rates lower than about 600 Hz) in the subglottal trachea, as the pressure is increased with vocal fold closing, and in the superglottal trachea as the pressure decreases with vocal fold opening. Also, propagating acoustic waves move the wall surfaces and propagate through the surrounding tissue, in response to the pressure cycle, reaching higher frequencies (up to 4 – 5 kHz) in both the subglottal and superglottal areas (Cheyne 2002).

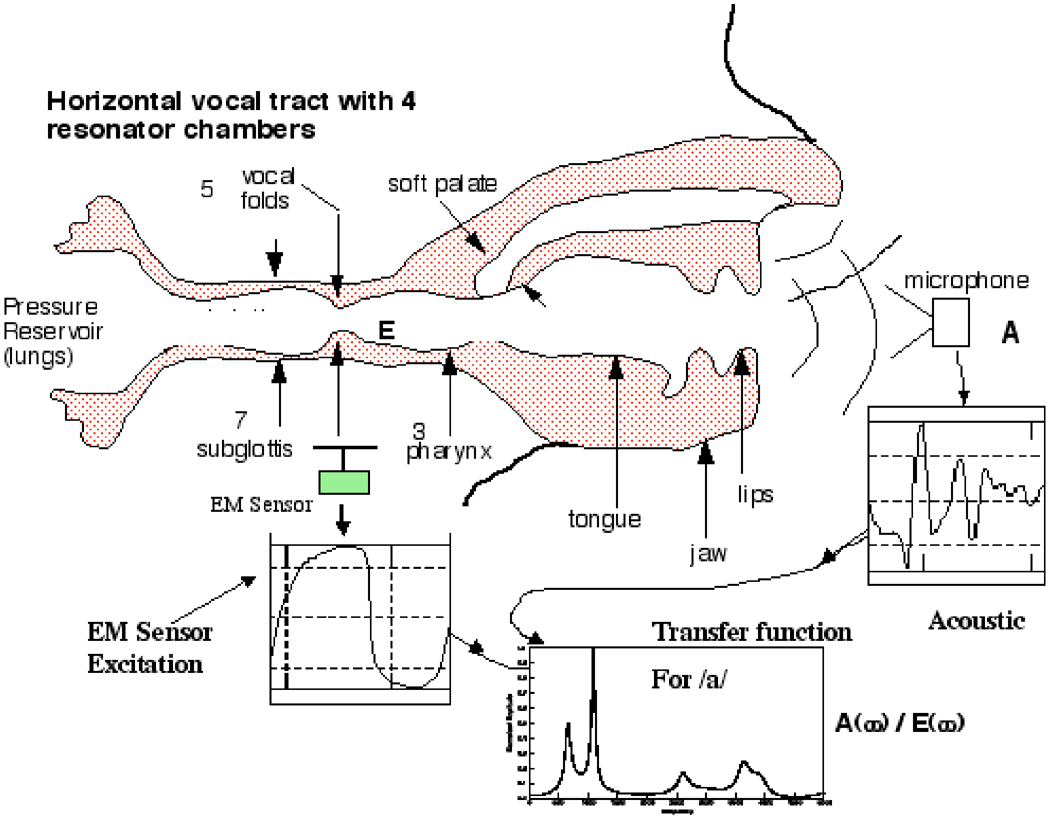


Fig. I-1: Schematic of vocal tract showing the main locations being measured for this report: Pharynx (superglottis) 3: vocal folds and glottal opening 5, and subglottis 7. Representative excitation, acoustic, and filter function data are shown.

Interferometric EM wave sensors (including laser-like sensors) can easily measure tissue motions at the one micron amplitude level and below. In addition, EM waves at frequencies below 3 GHz travel > 5 to 10 cm through human tissue before being excessively attenuated, which enables them to measure surface as well as internal vocal tract wall-surface and vocal-articulator motions during voiced speech (Holzrichter et al 2005). The data obtained by measuring the gottal opening using an EM sensor, can be associated with an air-flow excitation function (Flanagan 1965, Stevens 2000, Burnett 1999, Holzrichter et al 2005). Similarly, motions of internal vocal tract walls and overlying skin surfaces can be associated with air pressure pulsations (approximately 5 cm H_2O pressure changes), and hence an air pressure excitation function.

The purpose of this contract is to determine to what degree three different types of vocal tract tissues can be used to define pressure and/or volume excitations function for voiced speech characterization. A 2nd objective is to determine to what degree measurements of these three types of tissue are affected by loud, external noise. The three tissue types are vocal folds, vocal tract walls (at both super- and sub-glottal locations), and skin surface motions (overlying internal vocal tract walls). These were measured at nine locations on the head and neck of 9 subjects, speaking speech units of varying complexity, and with several measurements in the present of external tonal noise, up to 117 dB lin. Once the data from these measurements began to be understood, a method for obtaining a sufficiently good excitation functions for varying numerical speech characterizations (i.e., vocoding) applications was formulated, and is described herein.

II. Objectives of Darpa Contract with SNH Instruments:

Task 1: Estimation of voiced excitation functions using skin surface vibration measurements.

Task 2: Estimation of voiced excitation functions using tracheal wall vibration measurements using EM sensors.

Task 3: Estimation of voiced excitation functions using EM sensor measured vocal fold opening versus time.

Task 4: Model and summarize methods of voiced excitation determination and quality data from Tasks 1-3.

III. Experimental Approach:

Figure 2 shows 9 head and neck locations used for surface and internal vocal tract measurements. Surface measurements were made using a laser vibration interferometer (Polytec Instruments HLV-1000). Internal tissue motion measurements at the noted locations were made using EM sensors (e.g., SNH homodyne, Models G-011 and G-012, and a heterodyne sensors, model H-010).

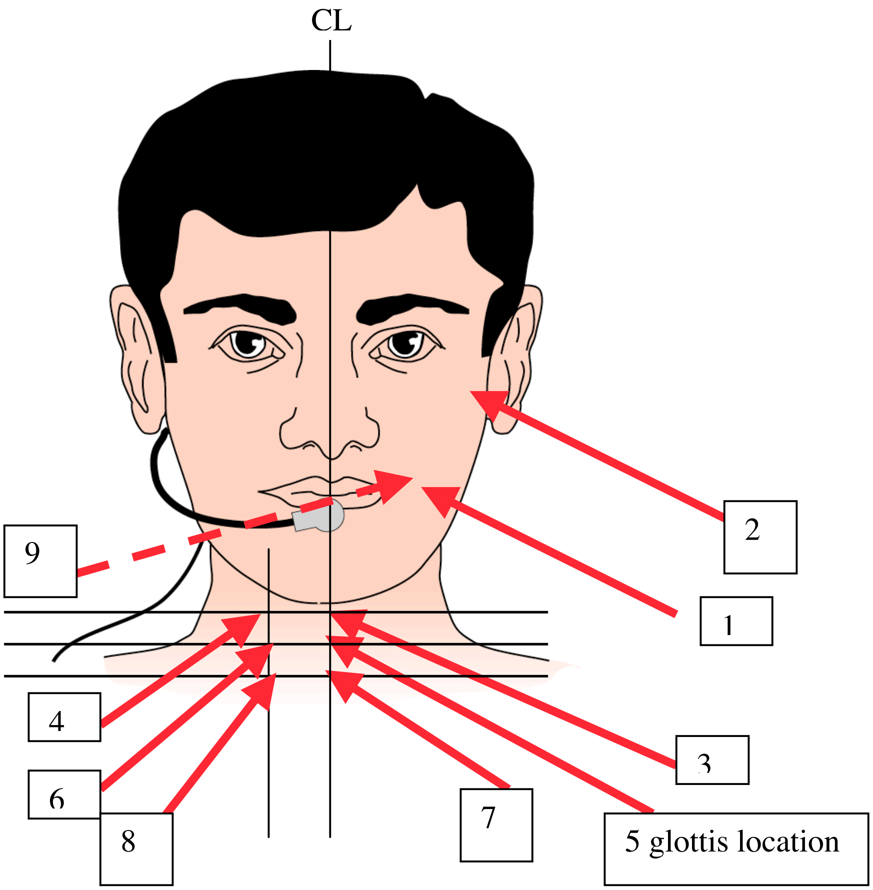
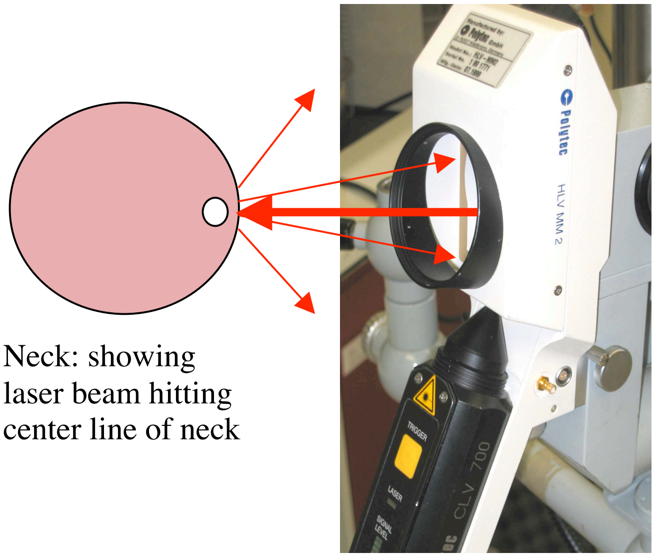


Fig. I-2: The locations of experimental data collection points on subjects’ faces. Location 1 is 2 cm from the ear canal toward the mouth, 2 is center cheek, 3 is on center line of head 3 cm above the glottis at the “super glottal” location, 5 is the glottis location, 7 is 3 cm below the glottis at the sub-glottal location. Locations 4, 6, and 8 are off to the side of the center line by 3 cm, at the super-glottal, glottal, and subglottal locations respectively. Location 9 indicates an experiment conducted with open lips through which the laser was shined onto the inner cheek wall, while speaking /ah/ . The boom mounted microphone, shown on the subject, was often replaced by a hand- or external mount-held dynamic microphone, or by a microphone connected into the oral cavity by an acoustic “feed tube”.

The laser interferometer sends a low power (about 1 mW) deep red laser beam toward one of nine tissue locations (see Fig. 2). The reflected laser light is slightly Doppler shifted by the skin velocity and is collected by the interferometer lens and mixed, inside the system, to obtain the Doppler shift, and hence the velocity of the target motion. A schematic of the system is shown in Fig. I-3 below.



Neck: showing laser beam hitting center line of neck

Laser Interferometer projecting a 1 mW red laser beam and then receiving scattered (and doppler shifted) light.

Figure I-3 Schematic of laser interferometer beam striking the surface of the subjects neck, on the center line. Locations 3, 5, and 7 are on the neck center line.

There were relatively few problems with this technique, once the reflective tape was applied to the target, and an experienced operator tracked the target with the laser. In addition, properly logging the data, keeping track of calibration settings, and polarities of the data (For example, skin expansion, due to increased internal air pressure, moves outward toward the laser and causes a negative signal from the interferometer).

The procedure for taking data was as follows: Typically a speaking subject (i.e., called “the speaker” in this report) would sit in a chair, an operator would direct the laser beam onto one of 8 external locations on the cheek or neck skin (see Fig. I-2). A small piece of reflecting tape would be applied to the speaker’s skin at the desired location, which enhanced the laser reflection by about 10x compared to direct skin reflection. The speaker would usually hold an EM sensor antenna against the skin, in front of the vocal folds at position 5. For data in Task 1 (skin surface measurements), a single output homodyne sensor (SNH G-010 which is GEMs-like in its configuration) was used for timing and glottal spectral data accumulation. A microphone would be mounted on a stand, at the desired distance from the mouth of the speaker. Data was collected at 44kHz, using SNH’s data collection system on one or two PCs, and stored in Matfiles for subsequent use. Procedures were similar to those described in detail in the papers by Burnett (1999), Ng et al (2001, 2002) and Holzricher et al (2003, 2005). The work was conducted under New England Institutional Review Board (NEIRB) study number: 04-113 .

EM sensor data for the present contract were taken in several ways, see sketch below. Prior data from Darpa, DOE, and NSF supported research were used as needed. The EM interferometric sensors operated at 2.4 GHz, generating 1 to 5 mW of power. They were used in three modes; 1) single ended homodyne (i.e, GEMs like); 2) I and Q homodyne mode (i.e., normal and phase quadrature); and 3) heterodyne mode giving amplitude and phase data from a vibrating target. In addition, sensors 2 and 3 had separate transmit and receive ports, which were used in two ways. Also, by using a circulator, a single waveguide antenna was used for transmitting and receiving EM waves. The single antenna configuration was used effectively in the super-glottal, glottal, and sub-glottal locations for localizing the sources of internal vibrations. The two-port configuration enabled us to use bistatic antenna configurations for measuring glottal opening by sending an EM wave through the glottal region, from one side to the other.

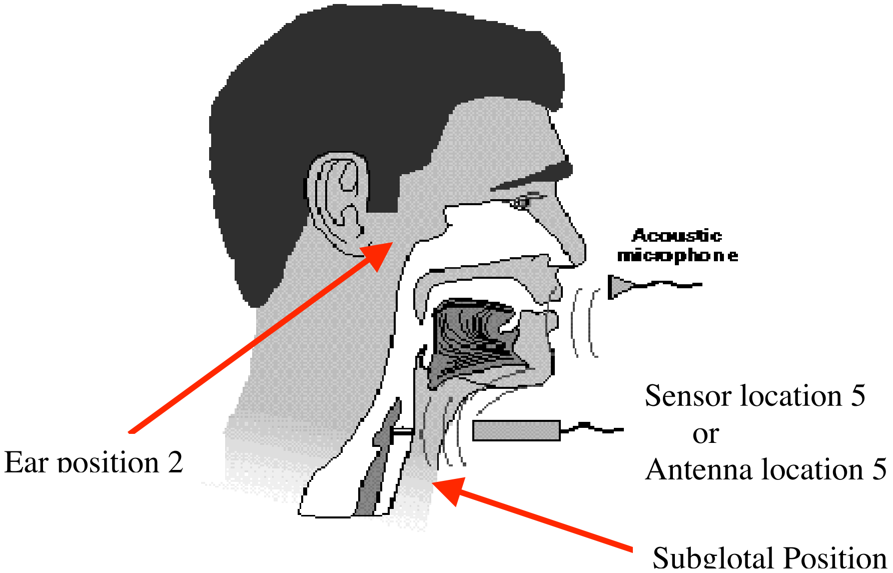
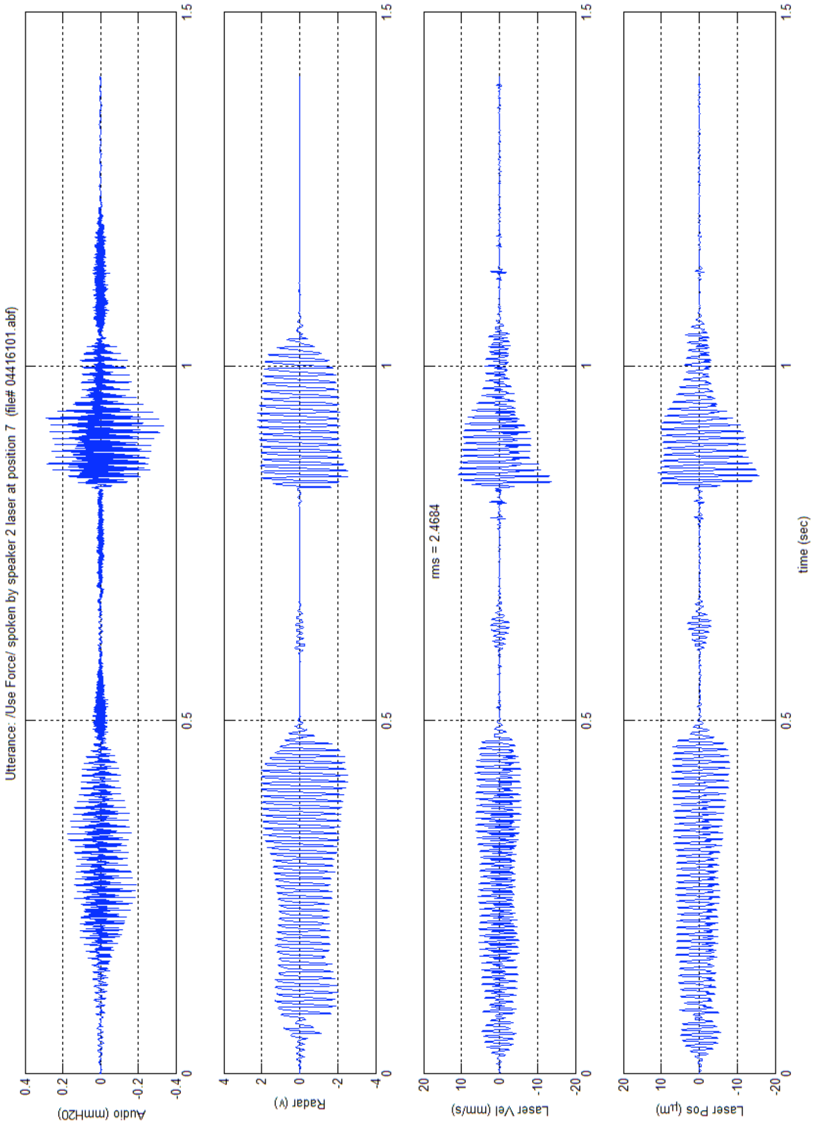


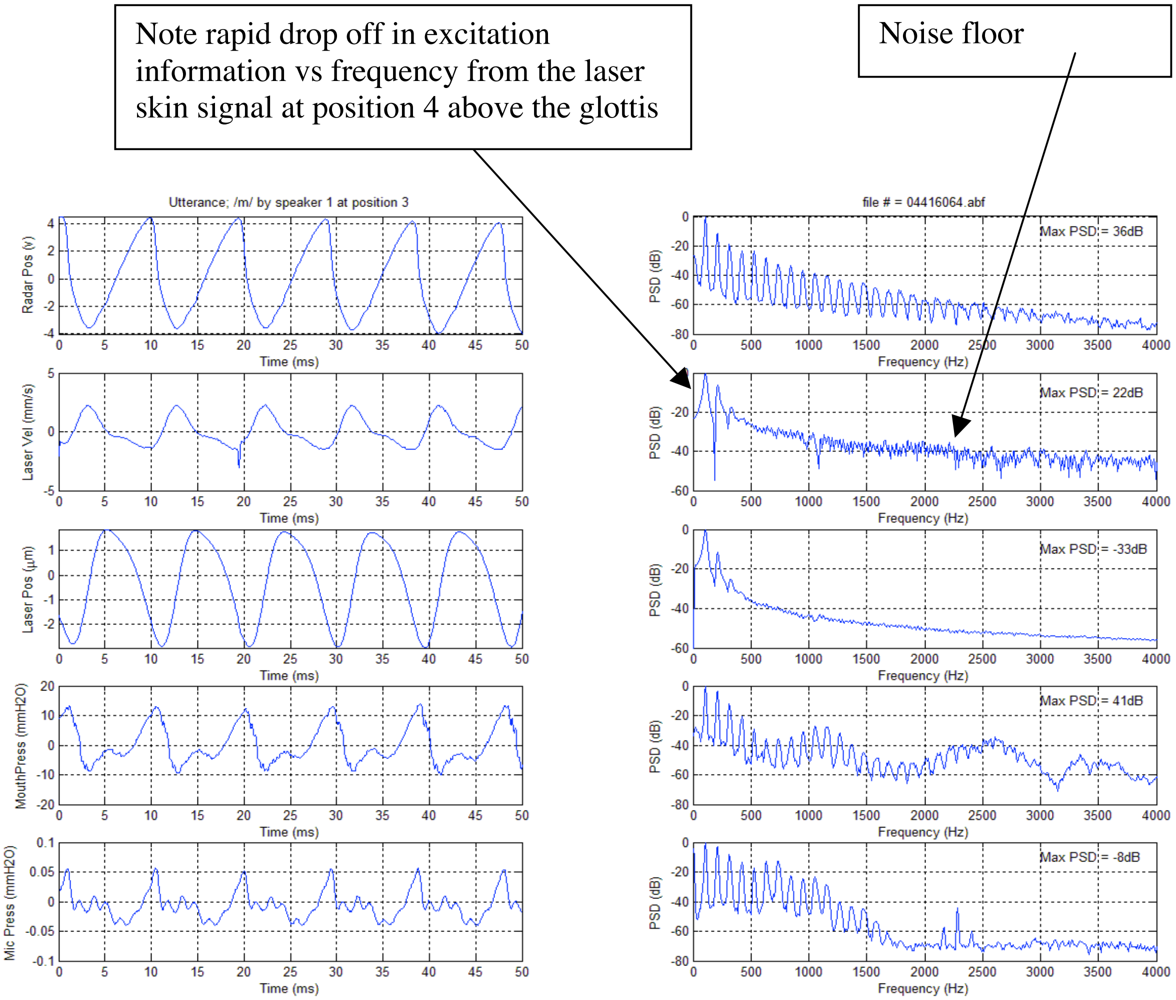
Fig. I-4 Typical arrangement of an EM sensor and antenna against the neck of a speaker at position 5, in front of the glottis. Also shown are two locations for laser sensed skin surface velocity, position 2 by the ear and position 7 on axis, below the glottis.

Figure I–5 below shows typical data from one subject taken with an acoustic microphone, a GEMs-like EM sensor located at the glottal region (position 5), and the laser interferometer velocity and the corresponding integrated velocity (i.e., position vs time) from position 7. The data are representative of signals from 6 speakers-male and female, speaking a sequence of phones -- /a/, /e/ , /o/ , and /u/ , then numbers 1, 2, 4, 9, 6, 16, 60, then letters a, b, d, g, k, m, t, and then the sentence “when all else fails use forc



One conclusion from Fig. I-5 above, is that the skin motion (traces 3 & 4) follows the radar glottal opening and closing well, and does not show the unvoiced segments of speech.

Fig. 6 below, shows male speaker voicing /m/ with the laser sensor at position 3, just above the glottis. One interesting aspect is the loss of excitation information versus frequency from this location above 600 Hz.



Note rapid drop off in excitation information vs frequency from the laser skin signal at position 4 above the glottis

Noise floor

IV. EM Microwave-interferometric Sensors and Antennas:

Pulsed EM wave (radar-like) sensors, i.e. GEMs, were first used in the mid-1990s to measure vocal fold motions, and shortly thereafter signal processing applications followed. (See early US patents by Holzrichter and Ng: 6,006,175 and 5,729,694). It was shown that EM waves from low power (approximately 1 mW) EM sensors, used in a dual-input/output homodyne interferometric mode, could measure the motions of the vocal folds, the jaw, tongue, lips , and other articulators as speech was being produced. It was recognized relatively early on, that a good measurement of the glottal opening (i.e., the vocal fold opening) versus time could provide information on speech airflow excitation functions, and a measurement of vocal tract wall expansion and contraction, due to pressure pulses from the vocal fold open/close cycle, could provide information on air pressure excitations. In addition, preliminary work was done to evaluate to what degree speech articulators, such as the pharynx, tongue, soft palate, jaw, and lip motions could be measured in order to provide articulator gesture information during human speech. It became clear that additional features were needed in EM sensors, optimized for speech applications, in order to extract full information from that which is available from the voice production system during speech. Additional features needed, compared to those available in early GEMs sensors, were extended target spectral response, higher gain, simultaneous phase and amplitude measurements, target location detection, antenna polarization control, mono- and bi-static measurement capability, and antennas for EM beam shape control.

SNH Instruments and Systems, the contractor, provided several of its EM sensor products for purposes of this contract’s measurements. Their new sensor family is based upon using low cost components, operating at 2.4GHz, to make EM wave interferometers for purposes of accurate motion measurement. These devices can make very accurate longitudinal position vs. time measurements, e.g., 1 micron positional changes measured each 100 microseconds, from relatively small target areas, of a few cm^2. The measuring modes used for the experiments herein included a single port homodyne sensor measuring both amplitude and phase (i.e., GEMs mode), a two port homodyne I and Q sensor for monostatic (i.e., single) or bistatic (i.e., double) antenna operation, and a heterodyne 2 port sensor for amplitude and phase measurements using one or two antennas. The sensors and antenna systems met or exceeded almost all of our goals by providing motion measuring bandwidth to 10 kHz, sufficient gain for measuring few cm^2 area targets, sufficient longitudinal sensitivity for micron tracheal-wall “ballooning”, E-field polarization control for polarization reflection experiments, monostatic output for use with a high gain waveguide antenna for target transverse location (plus or minus 1 cm), and bistatic output for using two antennas to measure tissue loss, to measure cross polarization reflection, and to test forward EM scattering transversely from one side of the vocal folds to the other. The one problem with our EM wave measurements is that until recently we were unable to obtain sufficient phase resolution, at the few-degrees of reflected phase shift accuracy, for a few desired experiments. This problem is being corrected and those experiments needing this information, namely confirmation of EM wave reflection simulations by Holzrichter et al (2005) and accurate target locating, are being completed. These delayed data do not affect the conclusions of this contract.

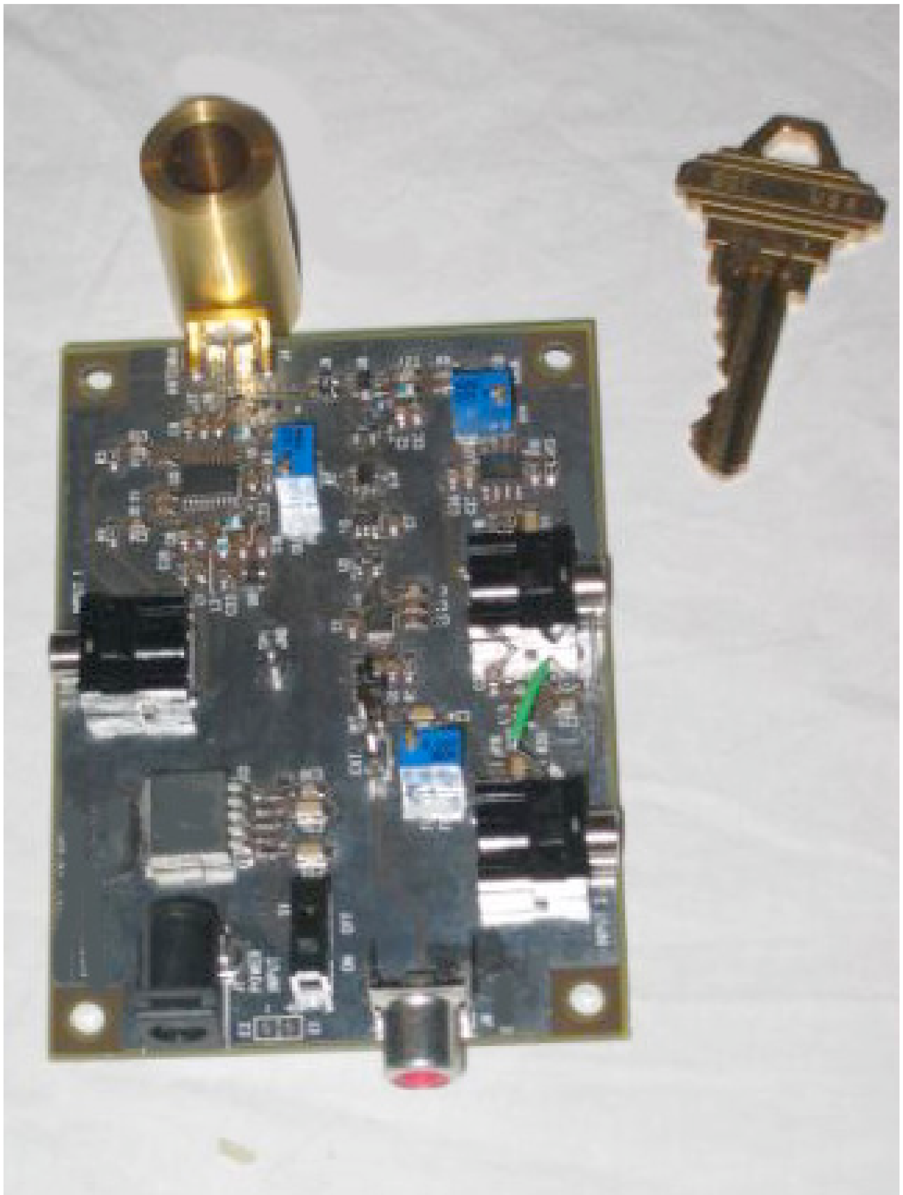


Fig. S-1   Single output homodyne sensor (GEMs-like) with a Small circular waveguide antenna, attached to the output SMA connector.

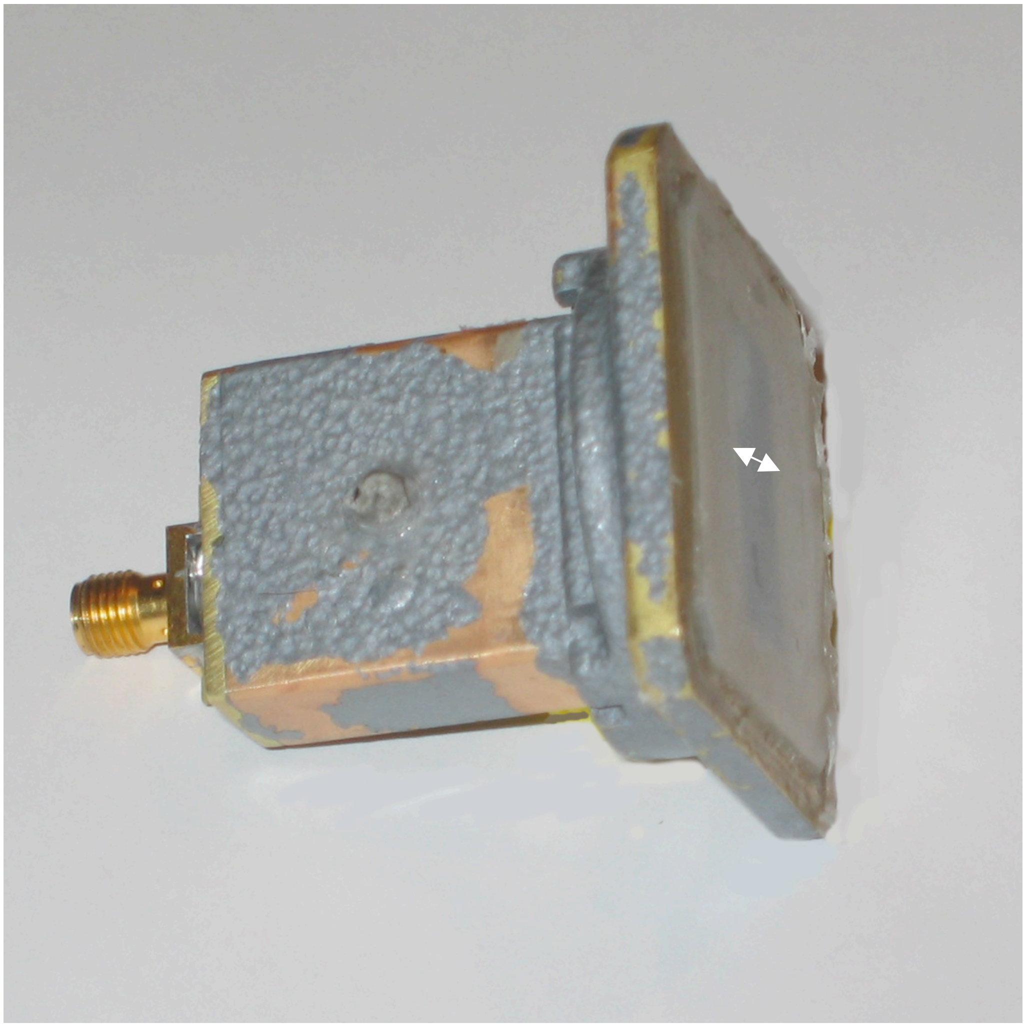


Fig S-2 A rectangular waveguide antenna matched to body tissue using acetone as the dielectric fluid. The arrow shows the direction of the E-field in the transmitted EM wave. This antenna can be used in monostatic mode as well as in bistatic mode, and it preserves polarization. The transverse dimension of the waveguide, also indicated by the double arrow under the plastic cover is about 0.5 inches. This dimension defines the narrowest EM wave beam-width. It was used to define target locations along the neck vertical dimension. Antenna produced by Cyberdynamics Corp, Palo Alto, under license from SNH Instruments LLC.

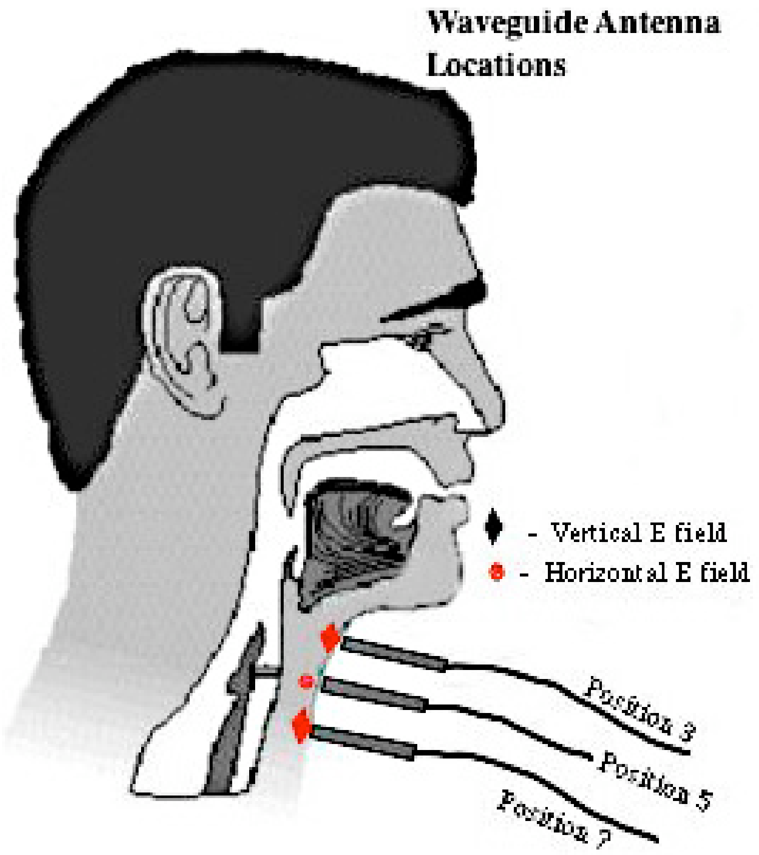


Fig. S-3 Illustration of locations of waveguide antennas for task 2 (which concerns positions 3 and 7, as well as 4 and 8) and task 3, concerning positions 5 & 6. At positions 3 and 7, the best signals were obtained using a vertically polarized E field; at position 5, a horizontally polarized E field worked best.

Figure S-3 shows 3 waveguide antenna positions for tracheal tube and glottal data collection. This antenna enabled sufficiently accurate target location that the EM sensed vibration data is unambiguous. In addition, data points 4,6, and 8 are to the side of the on-axis locations shown (e.g., positions 3,5, and 8). The side view resulted in low level, inferior signals to those on axis. The major reason for the lower quality EM sensor data from the side locations is that the curvature of the neck focuses the E field into the neck behind the trachea tube, reducing reflections from the trachea itself. A bistatic experiment was conducted (see below) using an index matching fluid (water) to planarize the skin surface allowing better targeting of the trachea by the wave guide antenna. In particular this improved the glottal signal amplitude by about 2X, or 3 dB. If it were desired to place an antenna to the side of the neck for “field-useable” configuration, it would be necessary to use an index matched “planarizing” or wedge-like layer to direct the EM wave to the targeted trachea tissue.

Bistatic transmit-receive experiments were conducted by using one waveguide antenna, of the type shown in Fig S-2, to transmit the EM wave and a 2nd to receive it. In this configuration we measure the transmission of 2.4 GHz through the skin and bone of a subject, observing about 3-5 dB per cm loss of transmitted signal. In addition, we measured the reflectivity of the tracheal tube versus E-field polarization direction relative to a vertical neck. We also measured the effect of E-field cross polarization (measuring de-polarization) by transmitting in one polarization and receiving in the cross direction. The results are that for the super glottal and subglottal locations, transmit and received E fields in the vertical (along tube axis) provide at least 2X more signal than horizontal or cross polarization. At the glottal location (position 5), horizontal field is superior by at least 4X in amplitude of reflection, giving the best reflection data. This occurs because this E-field polarization couples into the vocal folds through the trachea tube (consistent with simulations). A vertical E-field is not only lower in reflected signal, but the closure signatures are not seen. It is in fact similar to the EM sensor data from the trachea, shown in task-section T2 below, which is also best obtained with vertically polarized data.

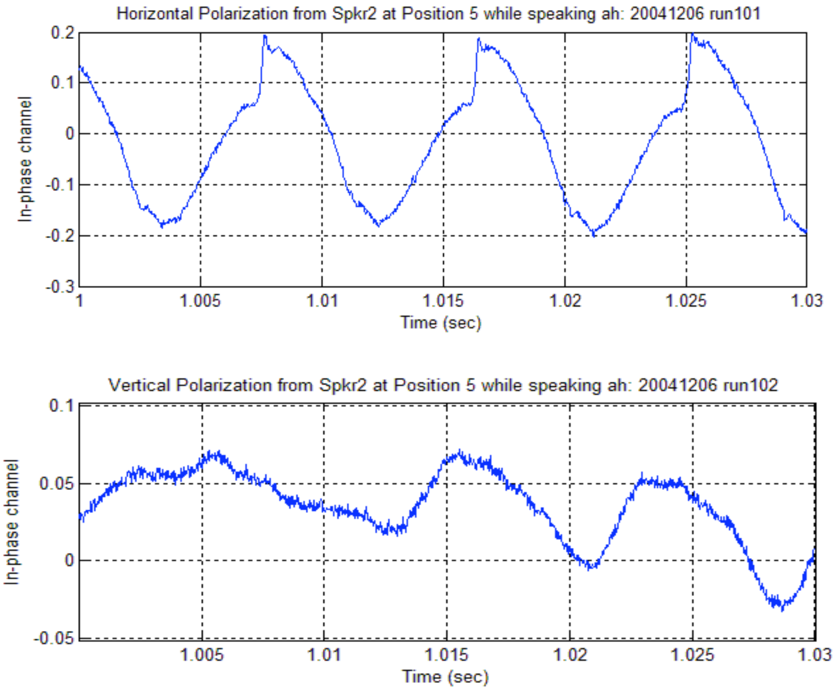


Fig. S-4 Examples of glottal data from position 5 in two polarizations using waveguide antenna. Note lack of high frequency detail in vertical polarized data, and the lower signal level, similar to tracheal tube data from position 3.

Other antenna configurations were tested. One particularly useful configuration was a “tuned” dipole antenna, whose antenna elements were chosen to best transmit and receive an EM wave into the high dielectric constant tissue surrounding the trachea and glottal region. This antenna had the advantage (or disadvantage) of measuring glottal signals from distances of 3 cm above and below the vocal fold locations. On the other hand, it was not able to measure tracheal tube motions directly because it had very poor localization ability, and it did radiate in the backward (out of skin) directions. Patch antennas were used in the original GEMs sensor by McEwan (1994) , in a closely spaced bistatic mode, and are commonly used for BlueTooth and WiFi communications systems these days. Patch antennas have the advantage of matching the plane of the skin well, but do not focus well (but better than a dipole) and they do radiate in all directions (forward and backward).

By using modern EM sensor designs, in an interferometric mode, with modern components now available for the 2.4 GHz communications revolution, and by using carefully designed antennas, we were able to obtain essentially all of the information that we proposed to obtain, with excellent quality. These data have enabled us to formulate a methodology for defining a voiced excitation function which we believe has sufficient spectral information to accomplish Darpa’s objectives for noise removal and narrow bandwidth vocoding.

V. External Noise:

For many applications of speech technologies, use in a high noise environment is desirable. However, most acoustic-only speech processing systems fail ungracefully in high ambient noise. It has been shown by many prior researchers that by fusing a noise-free signal from a second sensor –e.g., video, accelerometer, EM sensor, Electro-glottal graph – to the corresponding acoustic signal, significant improvements in recognition, speaker verification, and vocoding can be accomplished. The types of data being collected and examined for this report—skin vibration, internal vocal tract, and vocal folds—are susceptible to external noise to varying degrees. One objective of this contract was to measure the impact of loud external noise (> 110 dB) on the tissue systems being studied and on the sensors being used to obtain the data.

Experiments were conducted as illustrated in the Figure below:

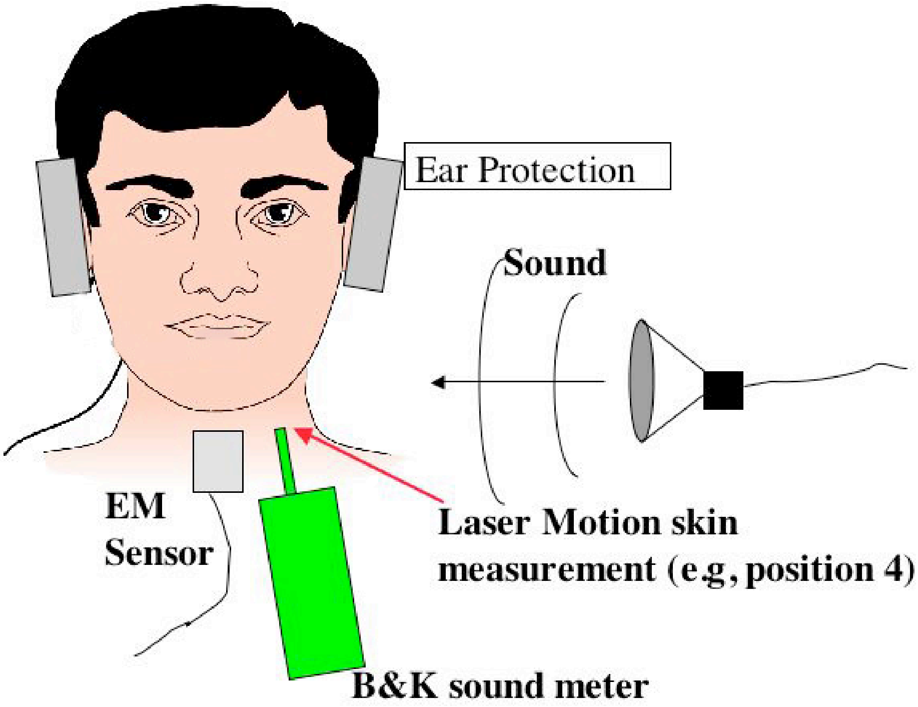
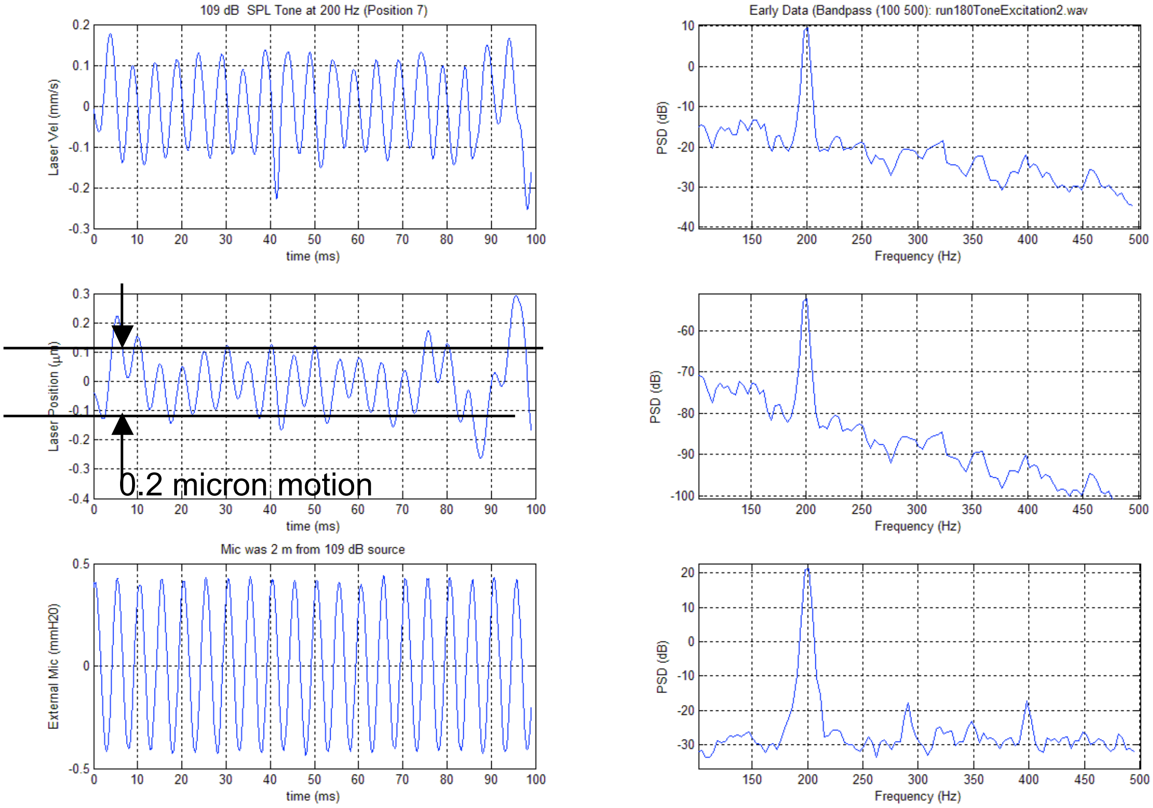


Figure N-1: Schematic for measuring cheek-skin and neck-skin vibrations induced by intense external tonal noise.

Loud tonal noise was generated by first recording 10 second intervals of 100, 200, 300, 400, 500, 800, 1000 and 2000 Hz sine waves on a CD. Then the signals from a CD player were amplified by a 400 watt, single channel solid state hifi amplifier, and were broadcast from an 8 inch speaker (in a wooden enclosure) onto the subject from a distance of about 1.5 ft away. The signal levels were set and measured with a Bruel & Kjaer type 1625 sound meter, set on the dB-linear scale (no spectral weighting was used). The probe was placed at the skin point being measured, then removed to enable the laser to reach the desired location (usually the cheek or the side of the neck at position 4). An auxiliary microphone (not shown) was used (at about 2M from the speaker) to record the noise signal electronically. In accordance with our IRB procedures, ear muffs were worn in addition to ear protecting inserts by the subjects; the subjects noted no discomfort due to sound. In addition, the instruments themselves — the EM sensors, laser velocity-meter, microphones, accelerometer, and the PC—were tested for response to noise. No unusual responses were noted. In particular, an accelerometer, when glued (e.g., with double sided tape) to the skin at position 4, registered only the skin motion due to the sound field (which is noticeable), and it added no spurious signals.

An example of raw data is shown below in Fig. N-2. On this figure the raw skin velocity versus time is shown in trace 1, with its PSD to the right, then the integrated velocity giving amplitude versus time is shown with its PSD (note that 0.2 micrometers of motion is measured), and finally the microphone output (located 2 Meters from the noise generating loudspeaker) is noted, with its PSD. By translating sound intensity in dB to air pressure in cm H_2O, 110dB of sound has a pressure of about 0.065cm H_2O. Comparing this data point to air pressure-versus tissue motions inside the oral cavity and inside the subglottal trachea (Holzrichter et al 2003, 2005), where 5cm H2O air pressure pulsations were measured to induced tissue motions of typically 10 microns, we would expect to see 0.13 microns of motion. We see 0.2 microns of motion in Fig. N-2. This estimate indicates that the measurements presented here are reasonably consistent with prior experiments.

We also note that at 800 Hz and above, there was essentially no tissue response (see sections below on Task 1 and Task 2 observations). Also, see the summary of tissue response versus tone, at 115 dB sound level, in Fig N-3 below. The observed roll-off in acoustic response versus frequency is consistent with our observations of skin and vocal tract wall tissue responses as measured by the laser velocity sensor and our EM sensors. A resonance (or anti-resonance) is noted at 500-600 Hz which is thought to be caused by the first vocal tract formant of the subject, called F1. Its influence on skin response is consistent with similar subglottal data taken by Cheyne (2002).



Trace 1 shows skin velocity versus time in response to the intense 200 Hz sound field, trace 2 shows the skin location versus time, and trace 3 shows a reference microphone.

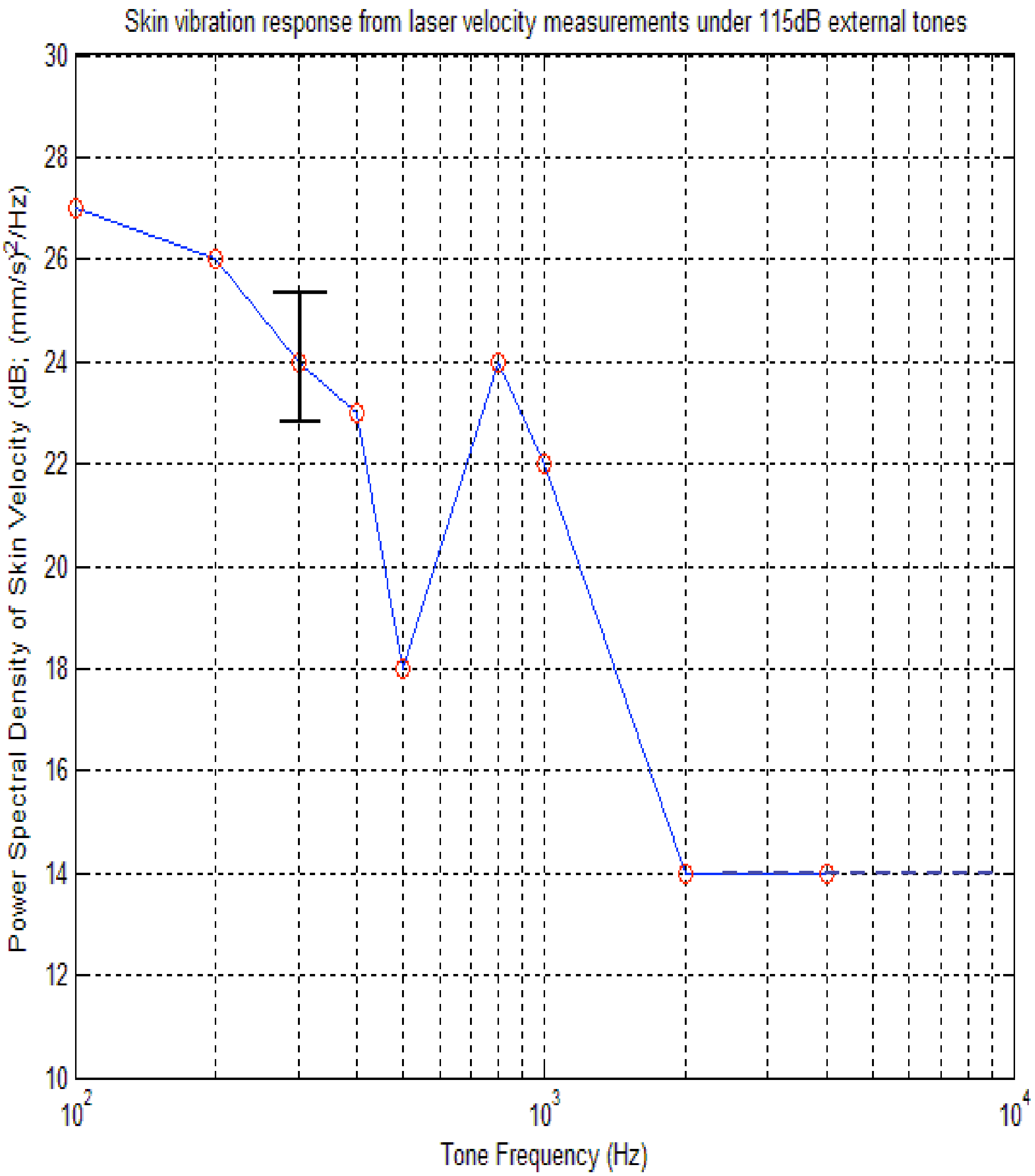


Fig N-3: Summarized skin response data from several experiments conducted using different tone frequencies at or near 115 dB lin. The error bar indicates approximate uncertainties due to level, skin, and data variations. Note the anti resonance at 400 Hz, and possible resonance at 800 Hz, estimated to be due to the internal F1 vocal tract resonance.

In conclusion, the data presented above, as well as other data that we have, and from reports on typical military noise levels and spectral qualities (see the ARCON sound disk) lead to the following conclusions.

1) For noise frequencies below about 800 Hz, narrow band noise induced skin position versus time motions, at intensities exceed 110 dB lin, are in the range of 0.04 to 0.4 microns, depending on the frequency. In contrast, measured skin positions vs time for speech induced motions are in the range of 1 to 5 microns. These are about 10 x larger than the narrow band noise induced skin position versus time signals. We add that noise, 10 db down, is easy to hear and will affect many automatic speech characterization systems.

2) For incoherent spectral noise, at intensities exceeding 110 Hz, which is “spread out” for example over 400 Hz, e.g., from 100 Hz to 500Hz, we would expect the same energy in the skin motion as for condition 1), but 10-100X lower amplitudes of vibration (depending upon the impulse response of the skin over the range of frequencies measured).

3) Different types of sensors measure different skin motion qualities. EM and optical interferometric sensors usually measure skin position versus time (x(t)). Doppler laser vibration detectors (e.g., the Polytek sensor user for tasks 1) measure velocity v(t), and accelerometers measure acceleration vs. time a(t). For single frequency motions:

$$a(t) = d/dt\ v(t) = d^2 / dt^2\ x(t) \quad \text{or} \quad a(t) = - \omega^2 x(t)$$

This formula illustrates why accelerometers are relatively insensitive to low frequency signals and are quite sensitive to high frequency signals. Position sensors, conversely, are more sensitive to low frequencies but become less sensitive to high frequency signals. This is one reason that the laser interferometer and EM sensor data is very clear at frequencies below 1 kHz and falls off rapidly above 1 kHz . Other reasons for limited high frequency sensor response to tissue vibrations has to do with the attenuation of high frequency acoustics by the vocal tract tissues and other nearby tissues themselves, and with the noise floors of presently used velocity and position measuring sensors.

4) Our data show that SNH EM interferometric sensors, measuring internal tissue motions of the vocal tract, are unaffected by external noise, up to 117 dB lin when their antenna is pressed gently against the skin, preventing the skin from vibrating. We also noted that one particular type of accelerometer that we tested, Knowles BU-3173, is insensitive to intense noise (when glued to skin and measuring the acceleration of the underlying skin). We also noted that the laser sensor was unaffected by noise levels near and above 110 dB lin.

VI. Contract Task Discussions: "VOICED EXCITATIONS"

The "Voiced Excitations Contract" has 4 tasks. The data and summaries of each of these tasks are discussed below. They are presented to satisfy the formal contract deliverables and the contract execution plan discussed with the contract monitors in May and June, 2004.

VI-1: Task-1 skin surface excitations.

VI-1A: Skin surface motions were measured in 8 external and one internal (oral cavity) location, see Fig. I-2 above. These data show that the skin and subcutaneous tissue structures overlying the vocal tract do not transmit high frequency information from the underlying air pressure excitations very well. These conclusions are indicated in the examples of laser data shown below and in the accompanying CD with other information included as part of this report. In addition, the transfer functions derived from oral cavity pressure function measurements to cheek skin motions support these observations (see below). Finally, work by Meltzner et al (2003) for purposes of developing an artificial larynx, show similar results. Fig T1-1 and T1-2 illustrate these assertions below:

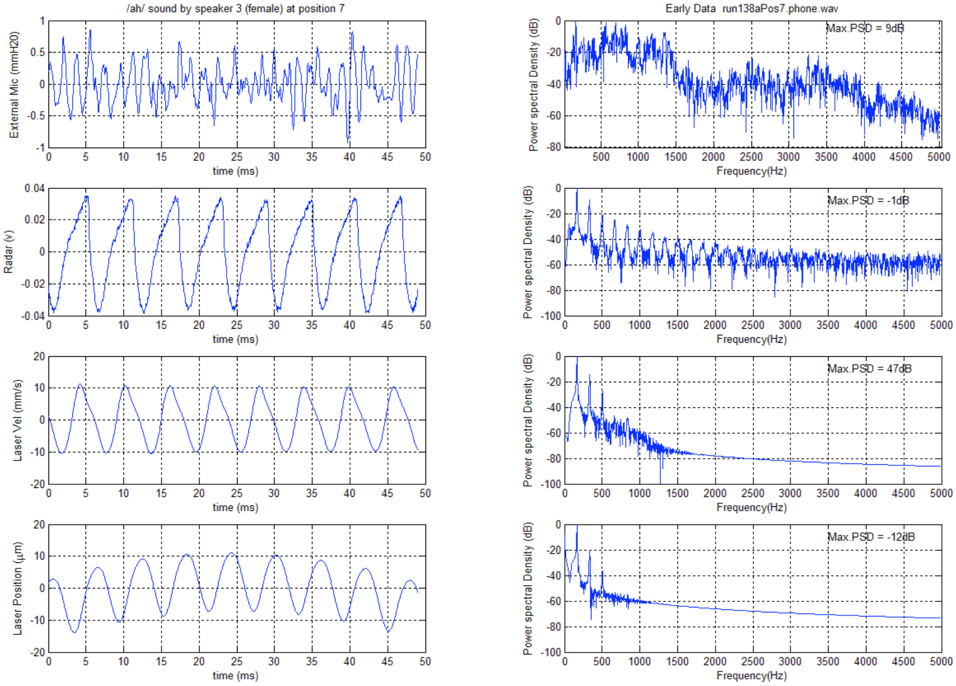


Fig. T1-1 : Acoustic, radar, laser velocity and position vs time and frequency for a male speaker voicing the sound /ah/ . Note the few harmonics in the laser PSD .

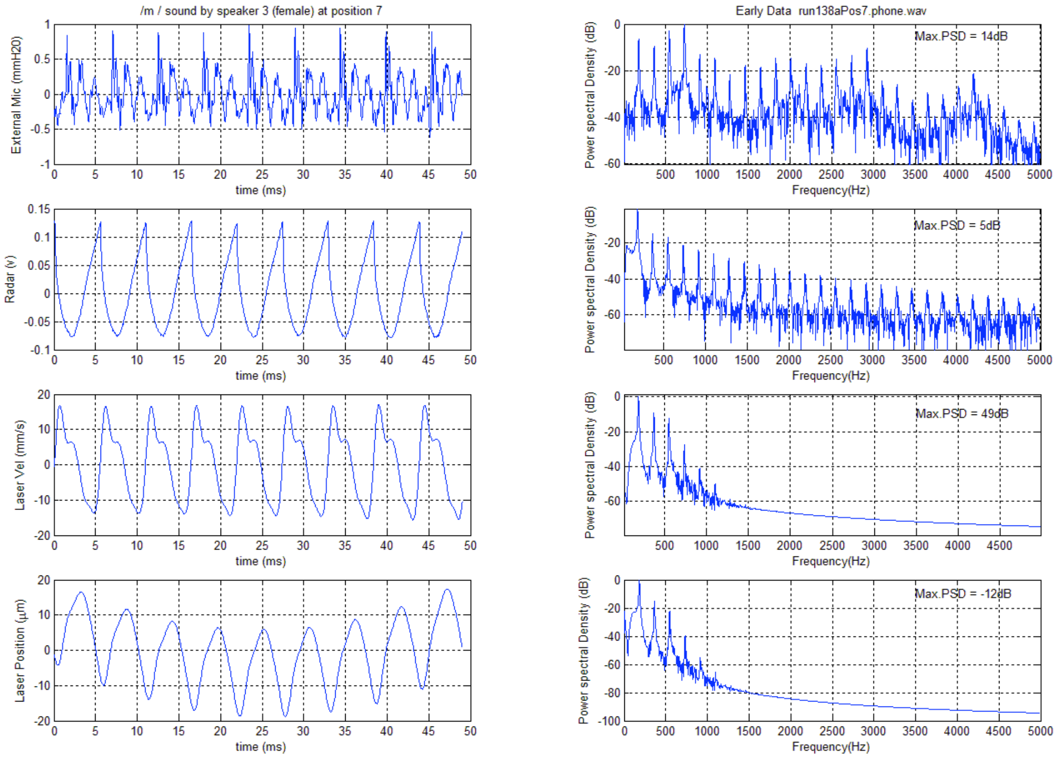
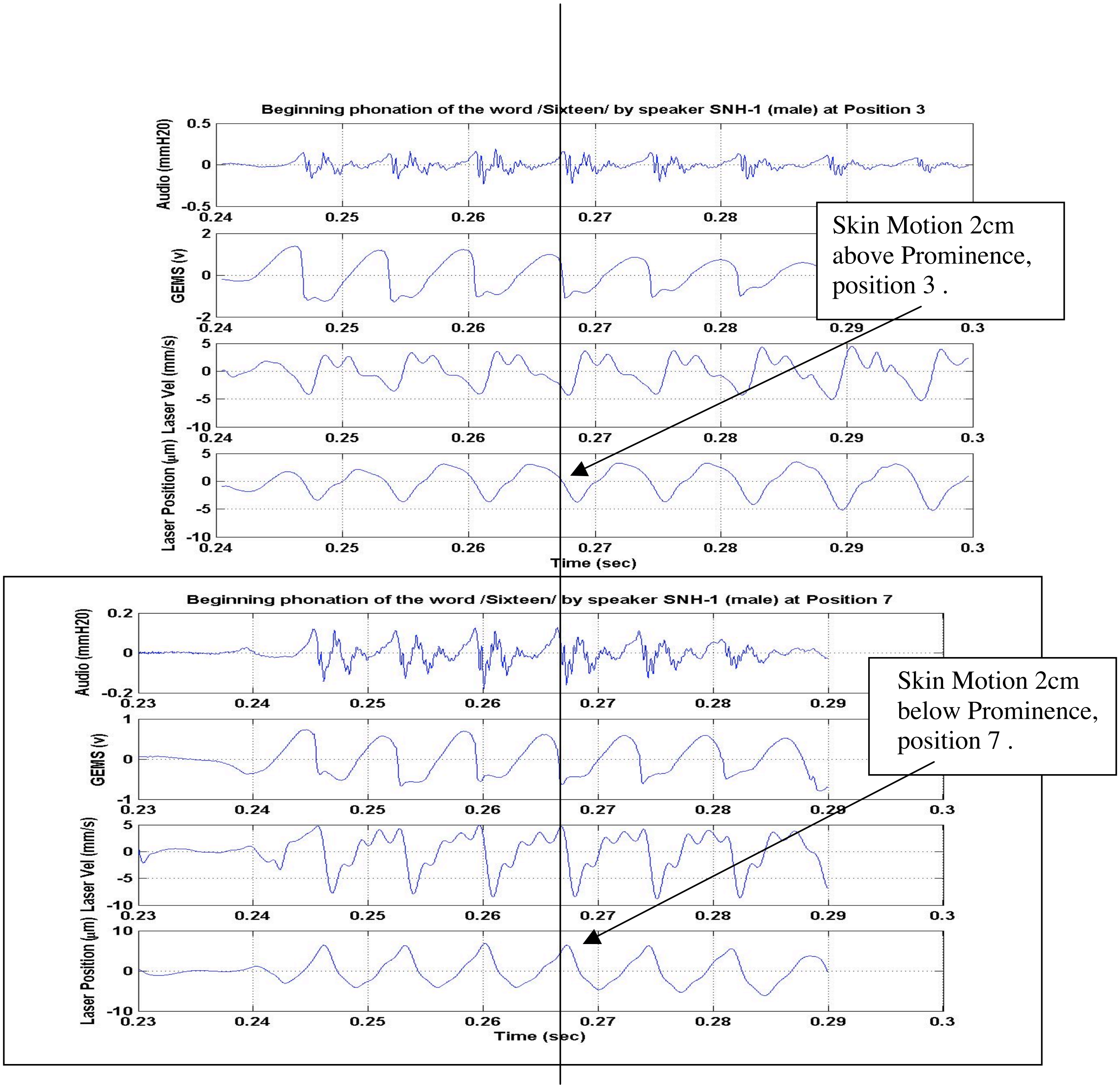


Fig. T1-2 : Acoustic, radar, laser velocity and position vs time and frequency. Male speaker voicing the sound /m/ .   Note the limited harmonics in the laser PSD, but also the fact that there is more spectral information than from the sound /ah/ above.

The data above show both the quality of the information available from skin motion detection and the challenge as to how to make use of the limited spectral content of these data. The objective is to develop a good excitation function for Darpa applications. These data show very reliable zero crossings of the skin motion yielding very good pitch period data. The few harmonics available provide sufficient information to define the fundamental glottal frequency and the glottal shape, which is largely determined by the drop-off versus frequency of the first few harmonic peaks and their bandwidths (after corrections for skin absorption, acoustic formants, and radar response functions). Dr. Ng, of our group, noted and shows in the summary of this report how this limited data can lead to an excellent excitation function.



Skin Motion 2cm above Prominence, position 3 .

Skin Motion 2cm below Prominence, position 7 .

Fig. T1-3 : Comparison between skin surface motion data taken at locations 3 (above the glottis) and 7 (below the glottis). The position vs time data for these should be opposite because the pressure pulses at these locations are opposite in sign.

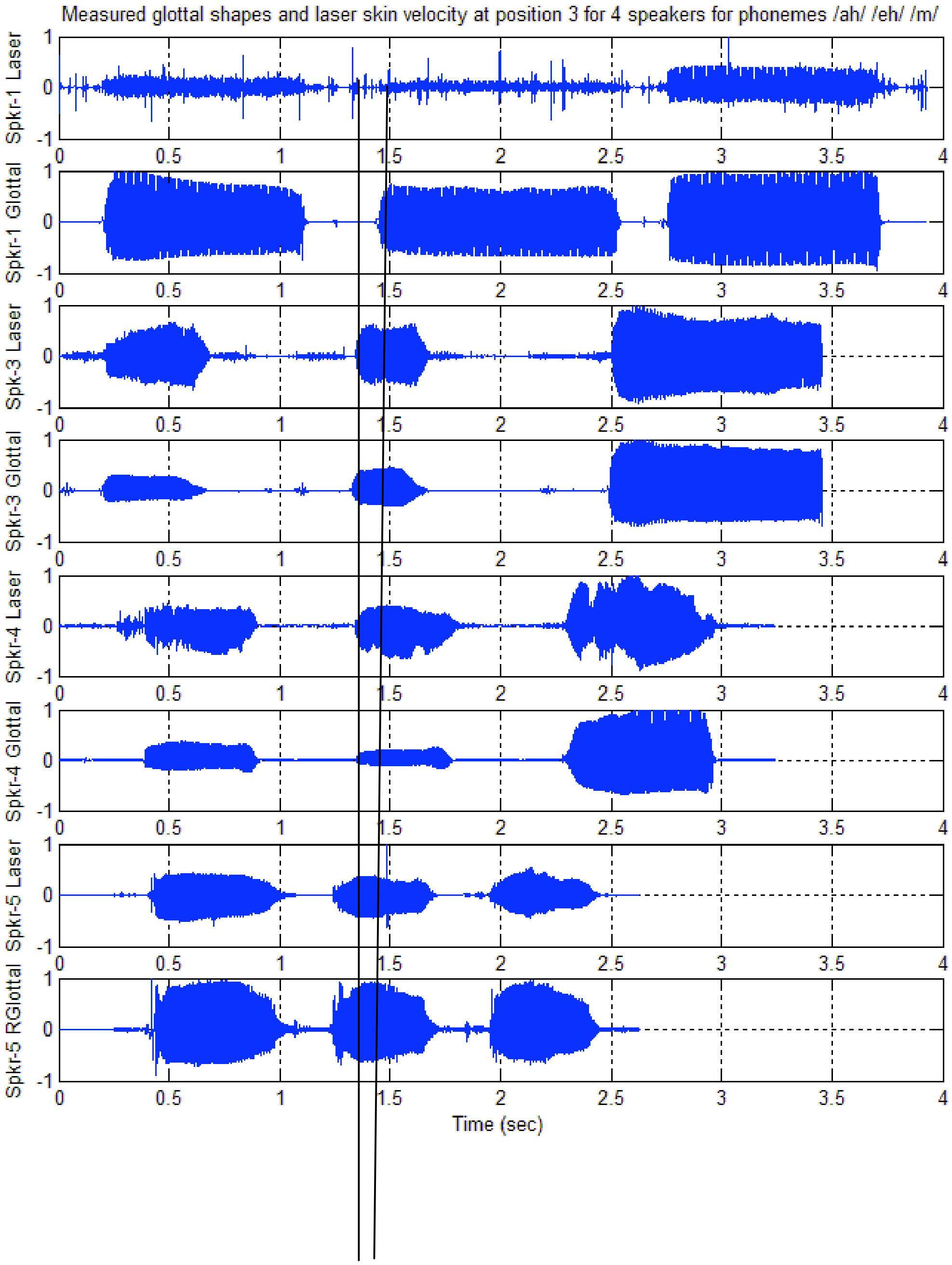


Fig T1-4A   Summary of skin velocity from position 3 (super glottal skin) and corresponding glottal radar signal for 4 speakers voicing the phones /a/ /e/ and /m/. The vertical lines define the region that is expanded in Fig T1-4b below.

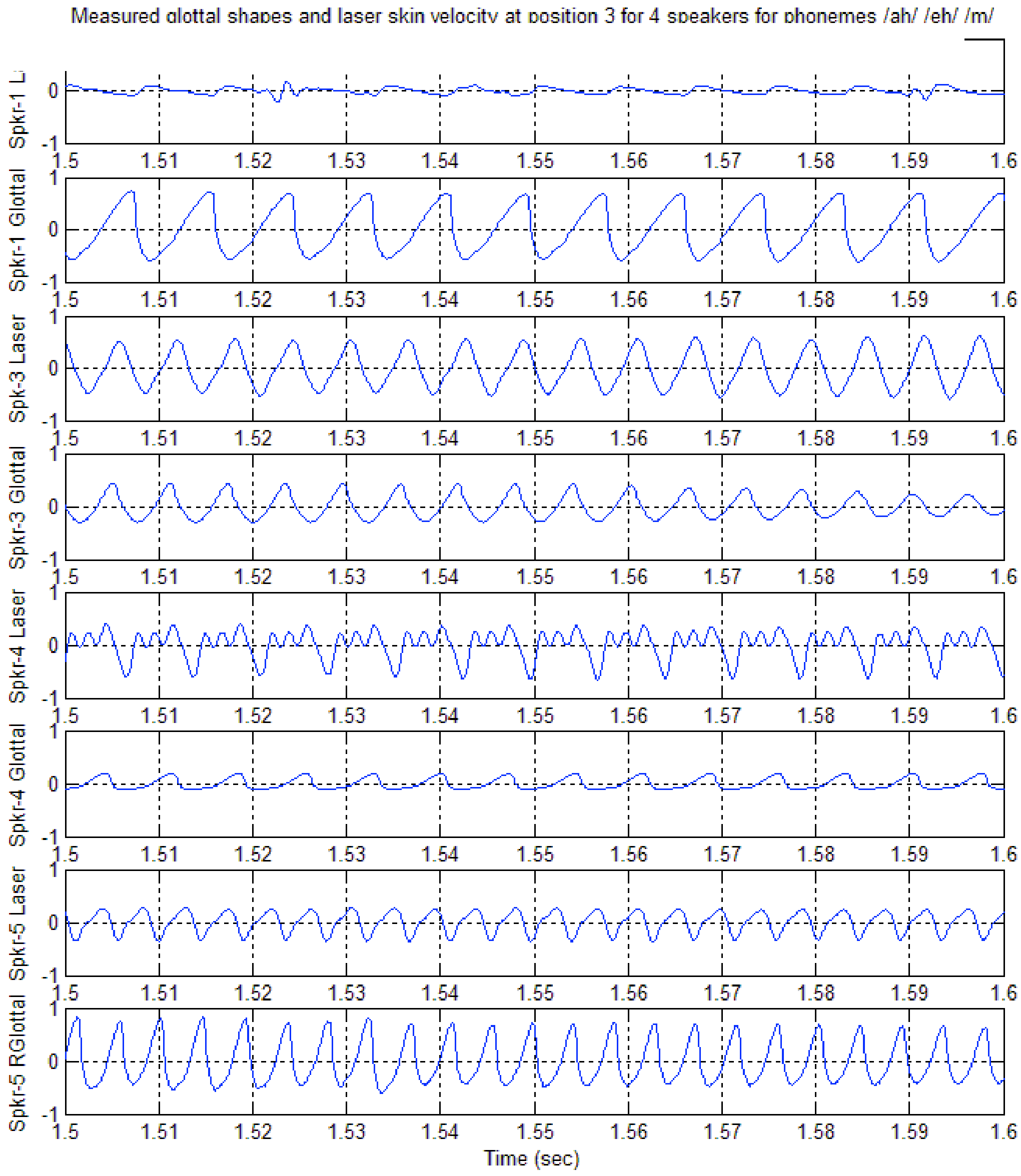


Fig. T1-4B: Expanded scale of position 3 (super-glottal skin) for 4 speakers speaking the sound /eh/ . Note the low amplitude for speaker 1, due to his weak phonation as indicated in the full data set shown in Fig. T1-4a above.

Additional survey data from 8 laser measuring locations are shown below:

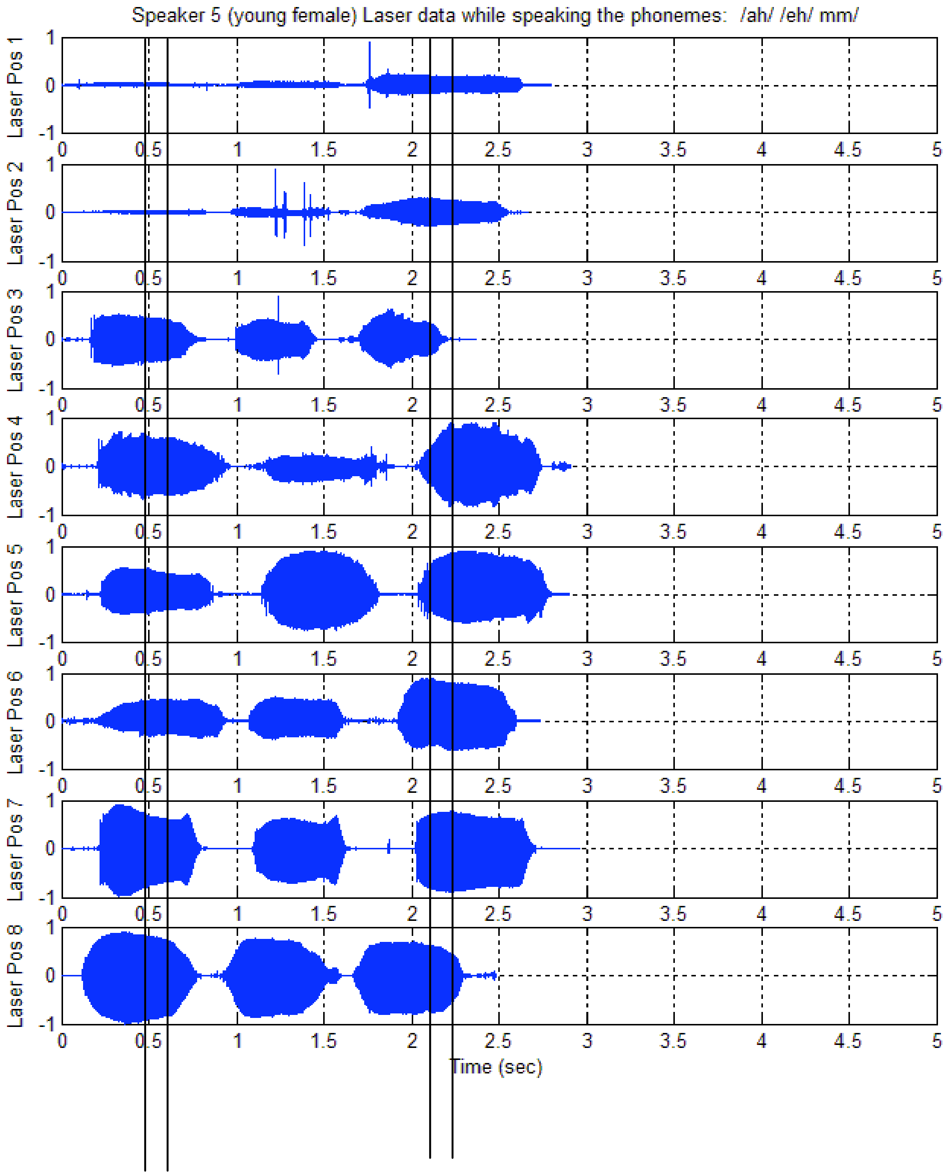


Fig. T1-5A: Summary laser skin velocity data from the 8 skin sampling locations from a younger female speaker. Note the weak signals from positions 1 (cheek) and 2 (ear) for open mouth sounds ah and eh, but the increased signal for closed mouth mm. The vertical lines show locations of expanded data, in the next two figures.

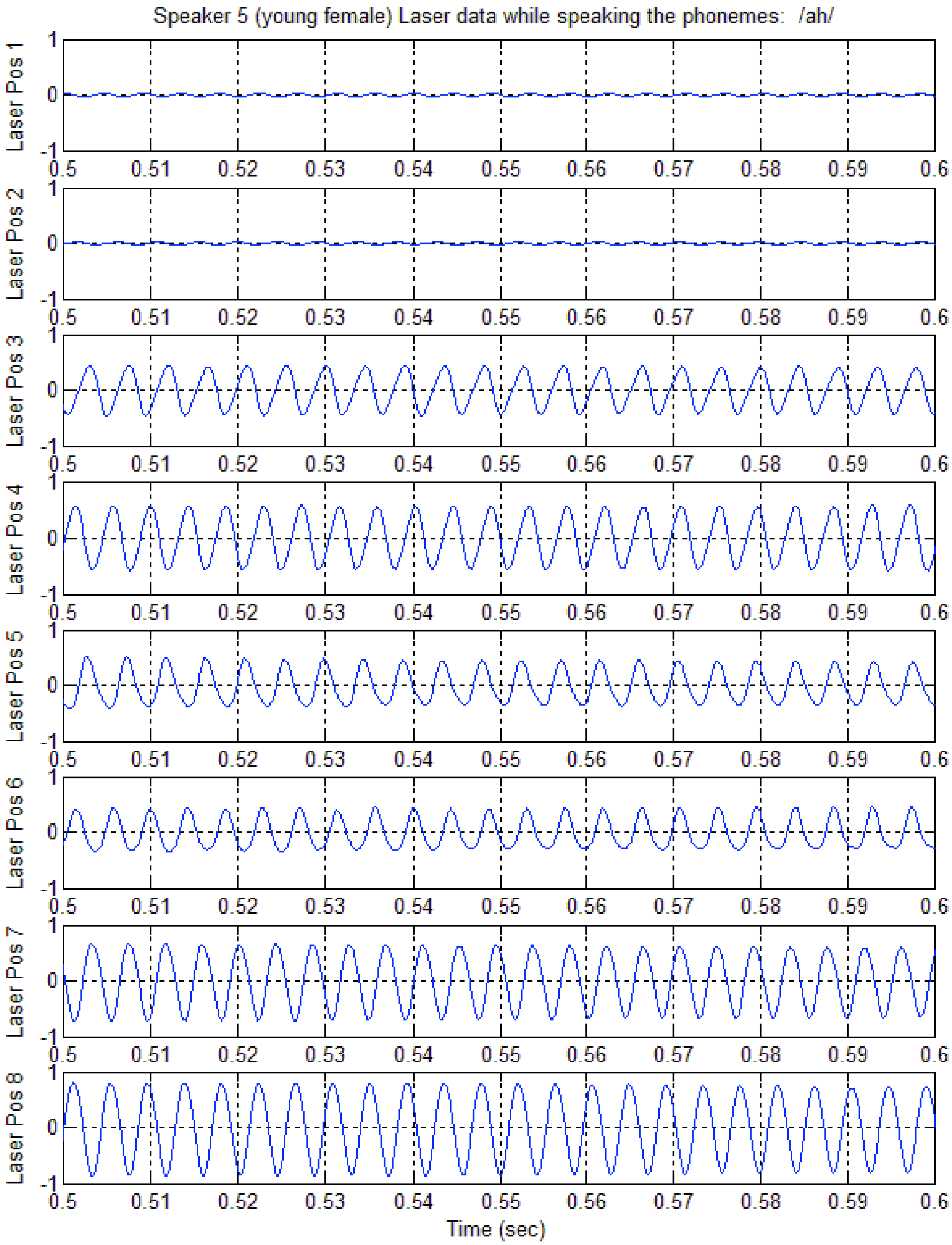


Fig. T1-5B : ah sound induced skin vibrations at the 8 skin measuring locations of a younger female speaker. Note weak skin motion signals from the open mouth (low oral-cavity pressure) associated with the sound ah .

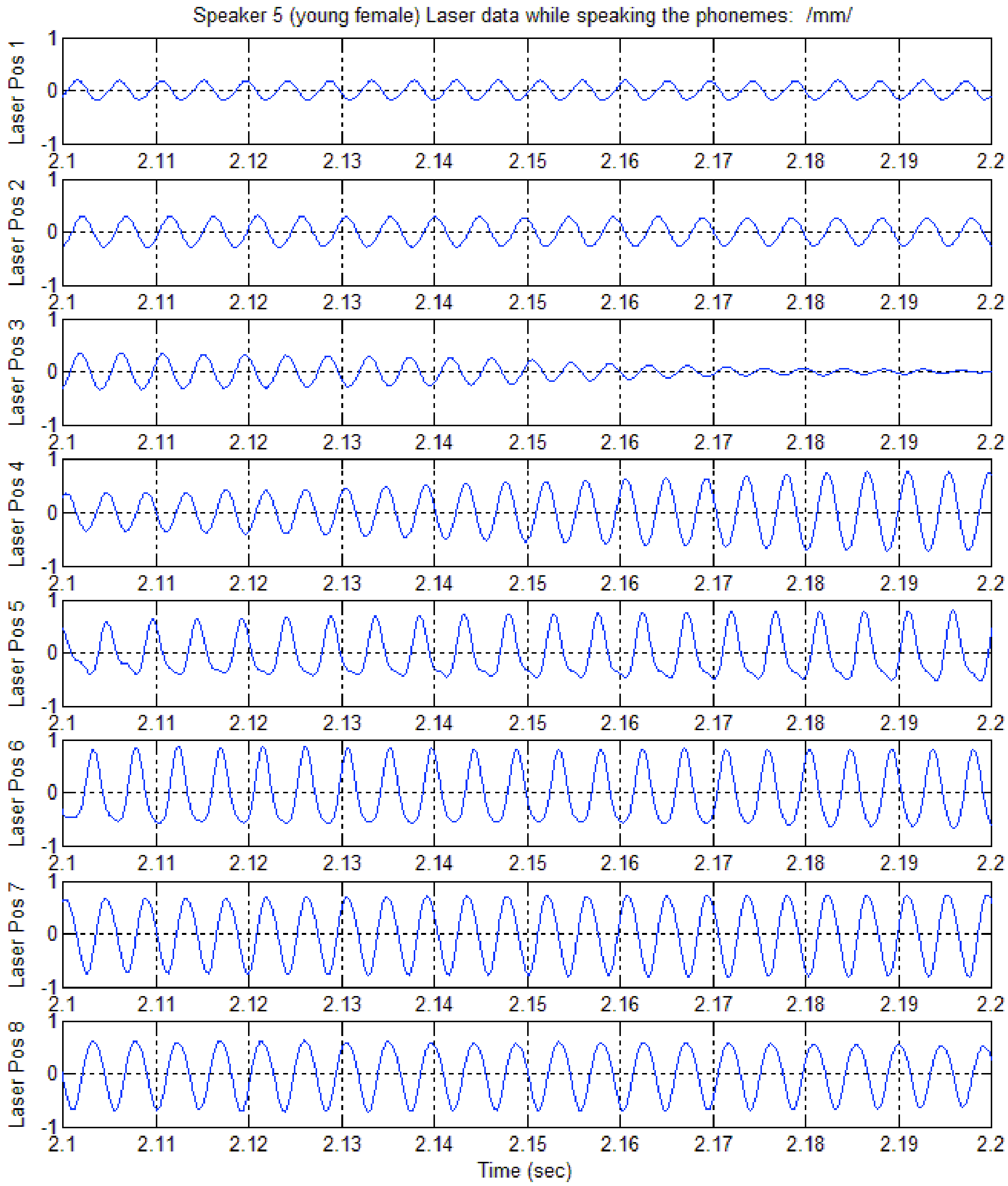


Fig. T1-5C: mm sound induced skin vibrations at the 8 skin measuring locations of a younger female speaker. Note stronger skin motion signals from the closed mouth (higher oral-cavity pressure, 5 cm H_2O) associated with the sound mm at positions 1 and 2 .

VI-1B: Transfer functions.

In order to better under stand the response function of the skin to internal pressure pulsations, we conducted an experiment in which we shined the laser Doppler instruments onto the cheek (position 1) while measuring the pressure of the oral cavity under several conditions. These data, shown below in Figures T1-6 A&B , illustrate the rapid fall off in internal oral cavity pressure to skin motion from the 100-200 Hz frequency range to 1000 kHz range.

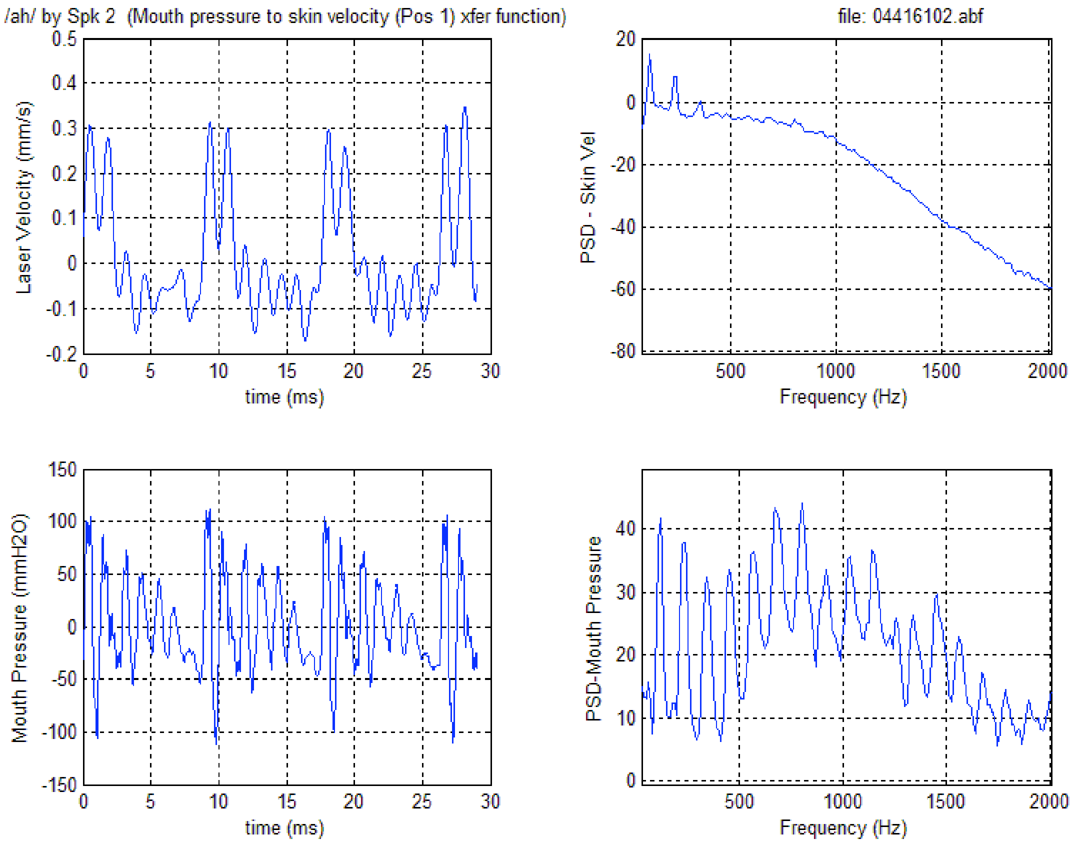
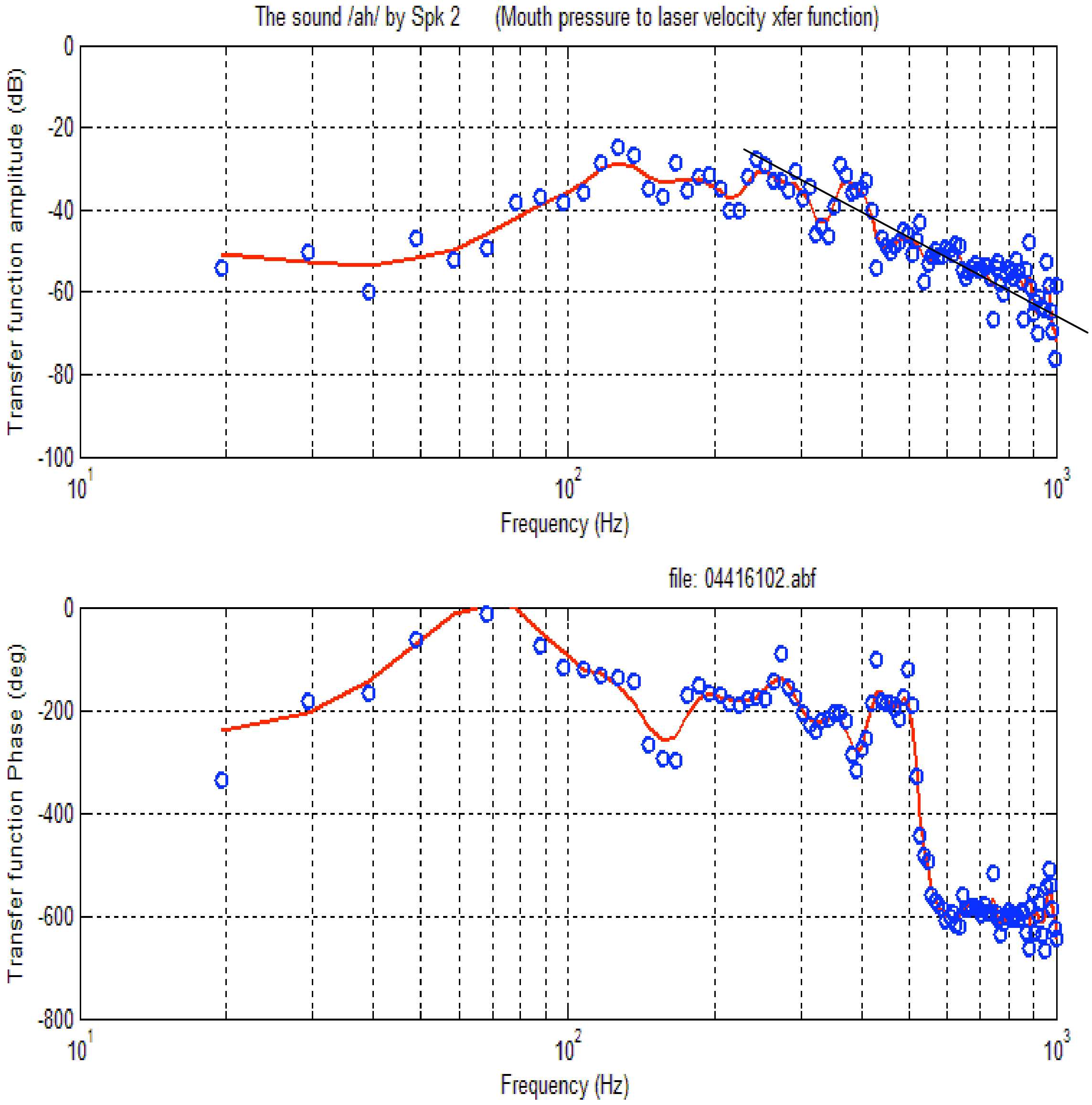


Fig. T1-6A: Raw cheek skin velocity as measured by the laser doppler instrument (position 1 on an older male) and the driving oral cavity pressure measured by having a “tube” convey the acoustic information from the oral cavity to an external calibrated microphone).



(
l
s

Fig. T1-6B: Transfer function from mouth pressure to check skin velocity. Note the –30dB rolloff in amplitude from 200 Hz to 1 kHz (which is –60 dB in power). Also, note large anti-resonance at 600 Hz in the phase plot.

Conclusions to Task 1:

The contractually proposed measurements and analyses for Task 1 have been accomplished. These tasks are enumerated in the formal list of 4 contractual Task-Deliverables and in the contract execution plan presented to the contract monitors in May and June 2004. For task 1, these deliverables are skin surface measurements of motion versus time for 8 external and one internal location, for six subjects with and without background noise, speaking a variety of phones, speech elements (e.g, letters and numbers), and a timit sentence “when all else fails use force”. Representative data and analyses meeting the commitments are shown primarily in the sections of this report labeled Section III Experimental configuration, Section V Noise and Section VI-1: Skin surface data.

In conclusion, skin surface vibration data can provide a great deal of information describing the presence of voiced speech and the qualities of the voiced excitation function. This is because it conveys accurate pitch information and sufficient information from the first few harmonics to identify a “ spectrally good” voiced excitation function. For the most accurate data containing the highest quality spectral information, corrections based upon the skin pressure response function, the acoustic resonances of the underlying vocal track segment, and the radar response functions must be carried out. However a great deal of information is available from the raw information itself. Finally, as discussed earlier in this report, external noise does play a role in the fidelity of skin response data, and it must be taken into account when using skin vibration information.

VI-2: Task-2 Internal vocal tract wall motion detection — Superglottal and Subglottal Trachea tisues:

The purpose of this task was to investigate to what degree tracheal tube vibration data can be used to construct a voiced excitation function. Figure S-3 in the sensor section of this report illustrates the experimental geometry. We find that the tracheal tube data carry more bandwidth information than surface skin data described above in Task 1, but they still roll off at about 1 kHz. Nevertheless, a good excitation function can be constructed based upon these data. In addition, these data are immune to external noise and are obtained relatively easily from throat-mounted sensors.

We have examined a wide variety of EM sensor data from the sub-glottal (position 7) and super-glottal (pharangeal) position 3. (Glottal data from position 5 is used for reference, see also Task 3 discussions below) The frontal views of the trachea tube yield sensor signals at the 10-50 mV levels (10X below the glottal signals of 0.5 to 1 volt), when the waveguide antennas are oriented to transmit and receive vertical E fields. Horizontal E field data are about 2 - 4x lower in amplitude than vertical. Using a bi-static antenna configuration, we obtained cross polarized field data at the 2 on-axis trachea tube locations, which are also strong enough to be useful, but are still weak and similar in amplitude to the horizontal field data.

It is very difficult to obtain good EM sensor signals from the side of the neck -- positions 4, 6, and 8 . (In contrast, skin vibration data is of good quality in these locations, see above in Task 1, Figures T1A,B,C). This is because most EM antennas can not aim the EM waves properly at the trachea walls, and hence do not receive good information. By pressing the antenna firmly into the neck tissue and aiming it toward the trachea, improved signals are obtained. It is possible that a specially designed neck-curvature compensating-wedge antenna can be constructed to relieve this problem. It is also important to note that almost all data taken previously, using wide angle antennas (dipole, patch, etc.), have not been capable of obtaining “clean” tracheal vibration data. The glottal reflectivity is so strong that it contributes signal strength through the aperatture (i.e., side lobes) of less directional antennas.

We find the subglottal tracheal wall data (position 7) sustains the excitation harmonics to about 1 kHz, and can be used to define an excitation function well. We believe, that once skin response, radar response, and subglottal resonances are accounted for (all relatively straightforward), these data can be very useful for voiced excitation function definition . However, the superglottal tracheal wall motion data falls off quite rapidly above 600 Hz (for both male and female speakers). We add that this is consistent with data we have obtained from air pressure measurements inside the oral cavity to cheek skin vibrations and from data by Meltzner (2003). We do not know why the superglottal region shows such fall off, but there are several explanations -- a zero in the transfer function at 500 Hz (which is near F1 which the first 1/4 vocal tract resonance), negative acoustic feedback reducing the apparent internal pressure, or the viscous damping of the tissues which become more muscle-fat-skin like, and less cartilage-like (i.e., the glottal and subglottal tissues have cartilage layers). In addition, see illustrated below, the superglottal region’s tissue configuration (positions 3 and 4) change depending upon the sound being spoken. In this region, the tongue slides up and down into the pharynx for different sounds (which can be detected by an appropriate low frequency sensitive EM sensor).

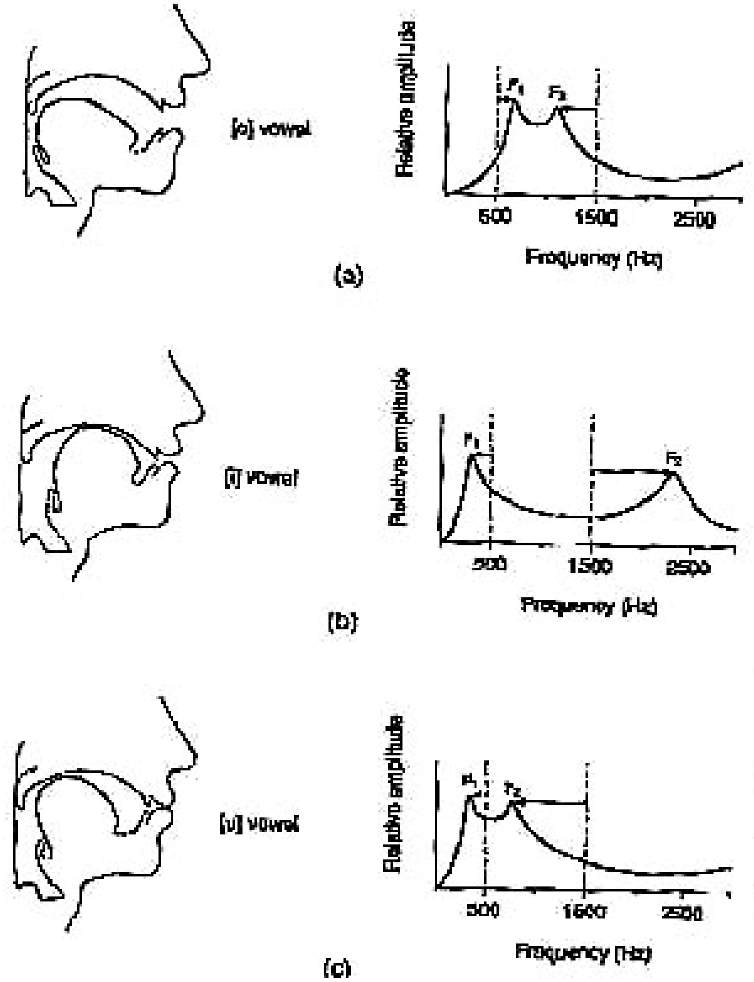


Fig T2-1: Illustrations of various sounds (Titze 1994) showing changes in vocal tract, which influence super glottal shapes and pressure levels. These changes, especially for closed mouth sounds such as /m/ , also influence subglottal pressures.

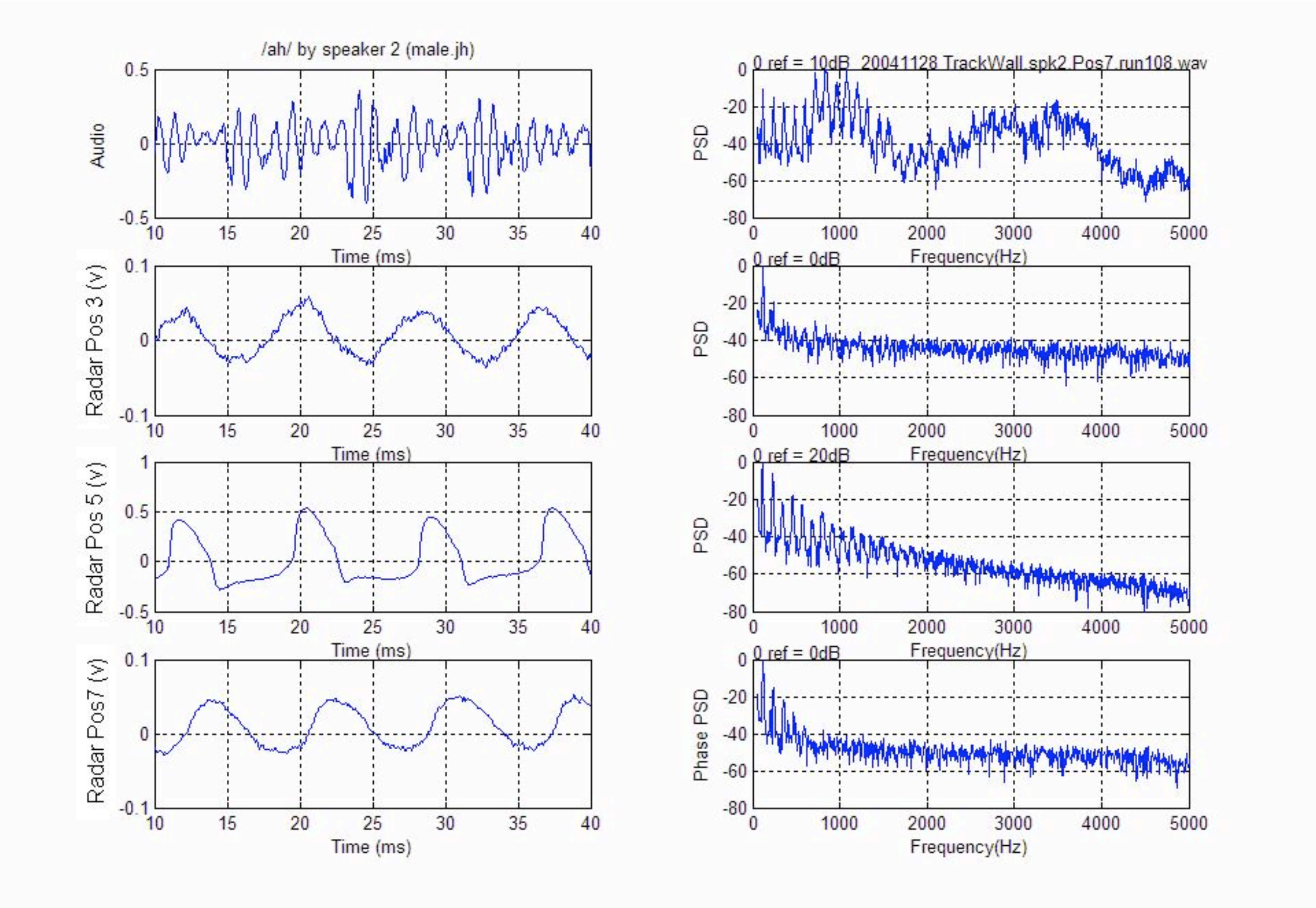


Fig T2-2: Male speaker #2 - EM sensed trachea data at 3 locations, super-glottal location 3, glottal location 5, and subglottal location 7. Note restricted harmonic structure in the PSDs for location 3 especially.

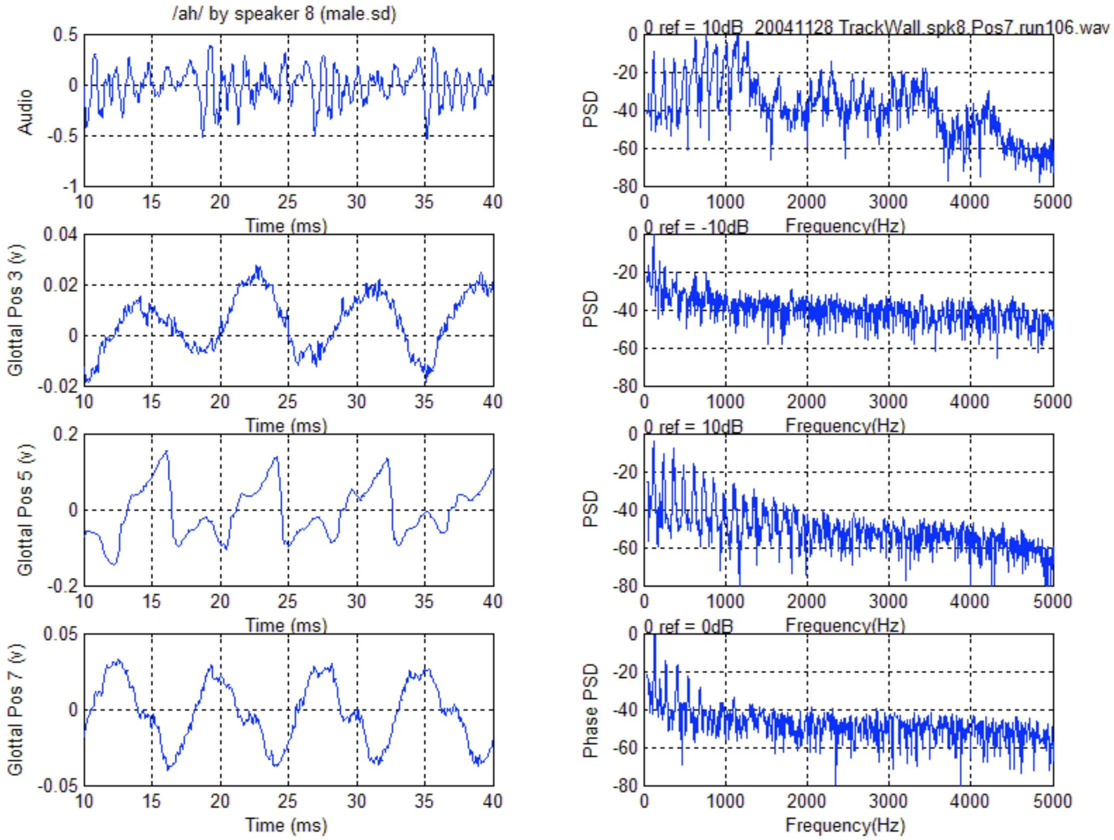


Fig T2-3: Male speaker #8 -EM sensed trachea data at 3 locations, super-glottal location 3, glottal location 5, and subglottal location 7. Note restricted harmonic structure in the PSDs for location 3 especially, but improved spectral content at position 7.

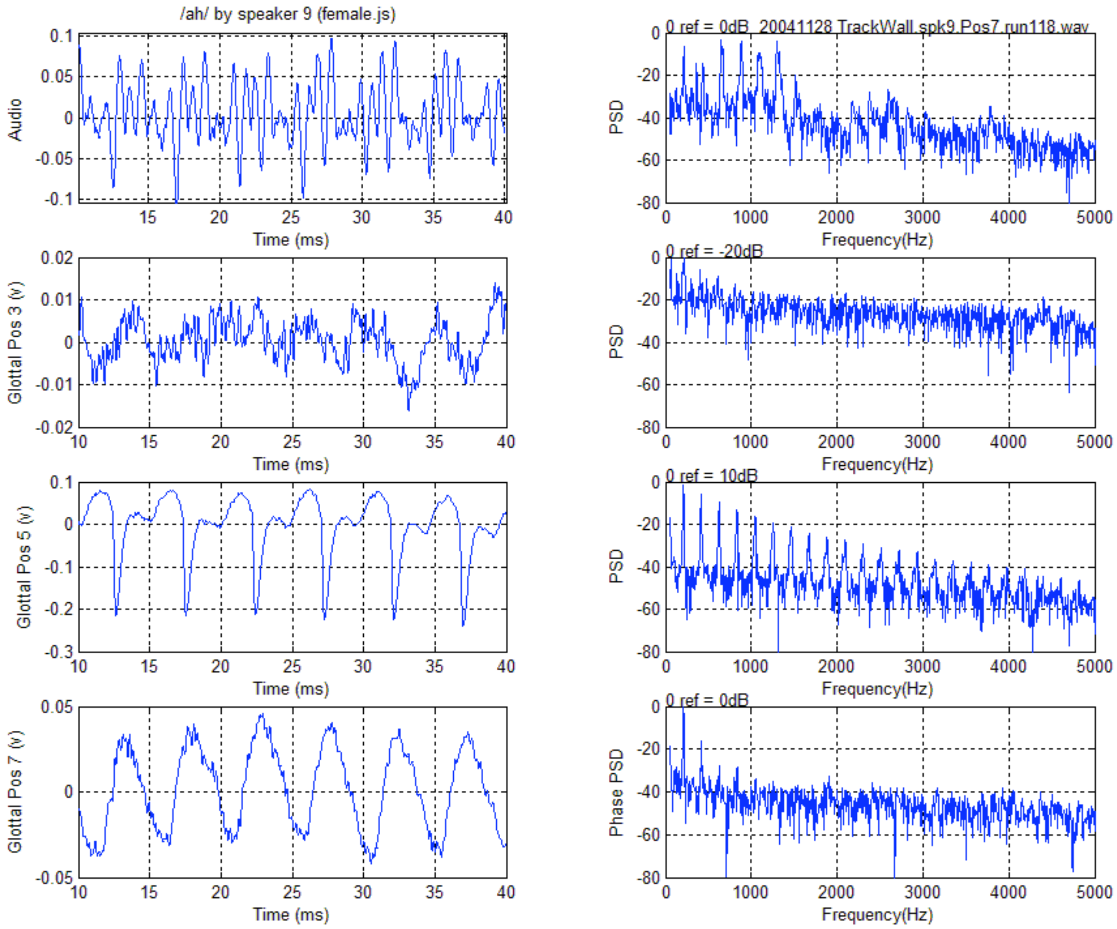


Fig T2-4: Middle age female speaker’s EM sensed trachea data at 3 locations, super-glottal location 3, glottal location 5, and subglottal location 7. Note restricted harmonic structure in the PSDs for location 3 especially, but also 7.

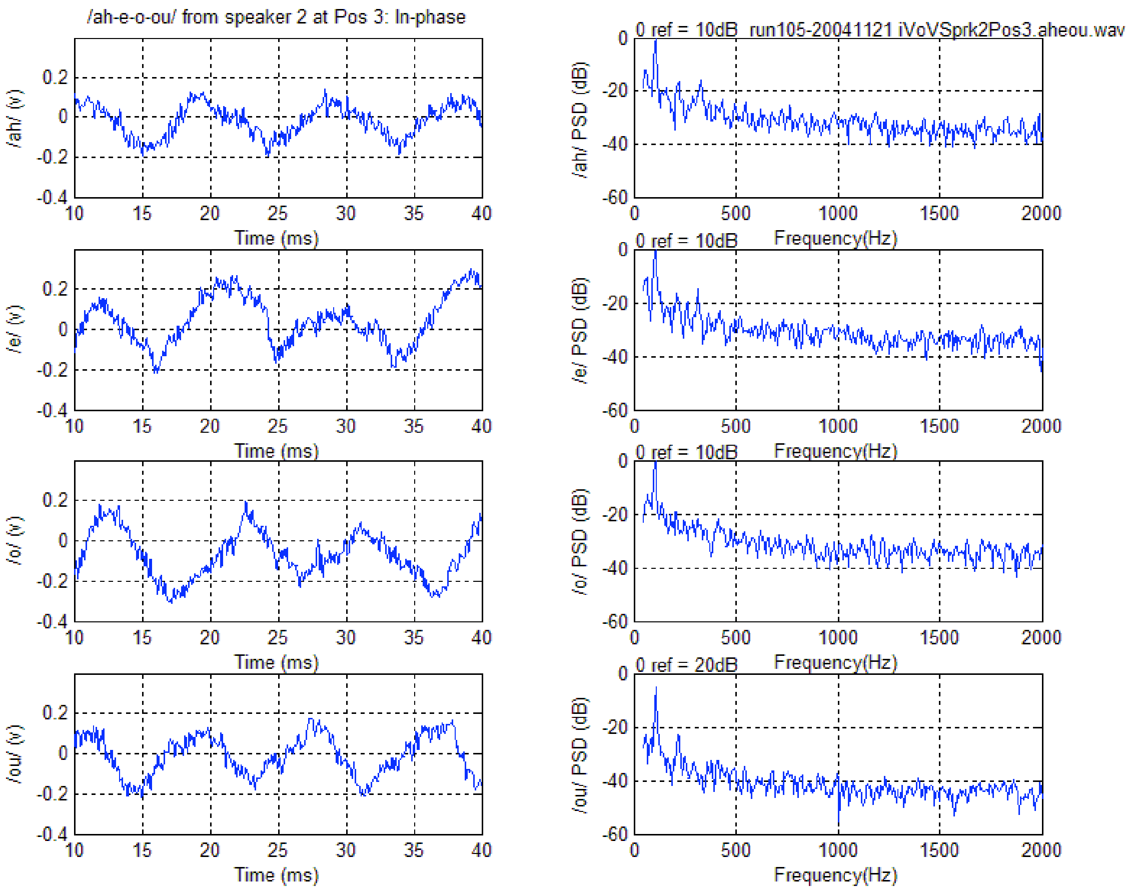


Fig. T2-5: Superglottal motions of trachea wall versus time. These show the relative strength of the signal versus the voiced sound being articulated. The changes are due primarily to varying superglottal pressure as the articulator configurations change and as the pharynx tissues change, see Fig. T2-1 above.

Conclusion to Task 2:

The contractually proposed measurements and analyses for Task 2 have also been accomplished. These deliverables are super-glottal and sub-glottal tracheal wall motions measured with a high sensitivity EM sensor and antenna, with a representative number of speakers (6 measured) and a representative number of sounds – 4 phones were used. Representative data and analyses meeting the commitments are shown primarily in the sections of this report labeled Section III Experimental configuration, Section IV EM Sensor and Antenna, and Section VI-2: Sub- and Super-glottal tracheal wall data.

On-axis EM-sensed tracheal-wall vibration information from the neck region, especially in the subglottal region 7, offers a very good opportunity to obtain voiced excitation data to construct a sufficiently good voiced excitation function to meet the needs of the Darpa “Vocoding in High Noise Program”. However other tissue locations, along the front and side of the neck can be used. Location 3 (super glottal) is also a robust location for tracheal tube measurements, but the vocal tract resonances and tissue reconfiguration due to pharynx and tongue motion, during articulation of different speech units, need to be considered in the interpretation. Positions on the side of the neck, locations 4, 6, and 8, pose problems for tracheal measurements using present antennas on EM sensors. If one of the side locations are desired for use, work should be done to develop proper antennas for conveying the transmitted and received EM waves to and from the trachea. The important finding is that the tracheal walls, especially the sub-glottal location, carries spectral information up to about 1 kHz, which can be used to define a “good enough” excitation function for most applications.

VI-3: Task-3 vocal fold data and excitations:

EM sensed vocal fold motions are well localized (within + 1cm vertically along the vertical neck axis) using the SNH sensors and the waveguide antennas. In addition, polarization measurements show that the horizontal E-field EM wave couples properly to the folds, and the vertical E field does not. These data are consistent with simulations (Holzrichter et al 2003 & 2005). Finally the SNH sensors have sufficient gain and dynamic range to obtain the needed spectral information for this report.

The glottal shape function as measured by EM sensors is a combination of many factors, but the rapid fall time is due to the glottal opening and closure. The detailed factors continue to be difficult to unfold in greater detail from raw EM reflection data, and it will require a several year program, probably with a university group to determine. In addition, it requires careful physiological work with groups such as those of Berry at UCLA who is taking careful pictures of glottal closure, and/or Puria at Stanford who is doing work on the mechanics of vocal tissues.. A complete understanding also requires improved EM sensors, with the ability to discriminate phase changes of a few degrees due to target reflections, but which are in the presence of several radians of overall phase change due to propagation through the sensor electronics, some air, and through other tissue structures.

Nevertheless, continued analysis of data from past and recent experiments show that the important spectral aspects of EM sensed glottal data from position 5 are very useful and do enable a very good voiced excitation function to be created. The key pieces of information needed for obtaining a good excitation function are pitch period, glottal pulse width, and glottal close time with sufficient spectral resolution and dynamic range. For practical use of these data, methods of numerically storing and retrieving such data in an efficient way is needed. We believe that we have now understand glottal excitation data from position 5, as well as data from other neck locations such as the super and sub glottal regions, sufficiently well that it can be used to meet the needs of Darpa’s and other programs. Details of the models and applications are discussed in Task 4 below.

Examples of glottal data for this report are shown below. They are displayed to emphasize aspects useful for speech characterization.

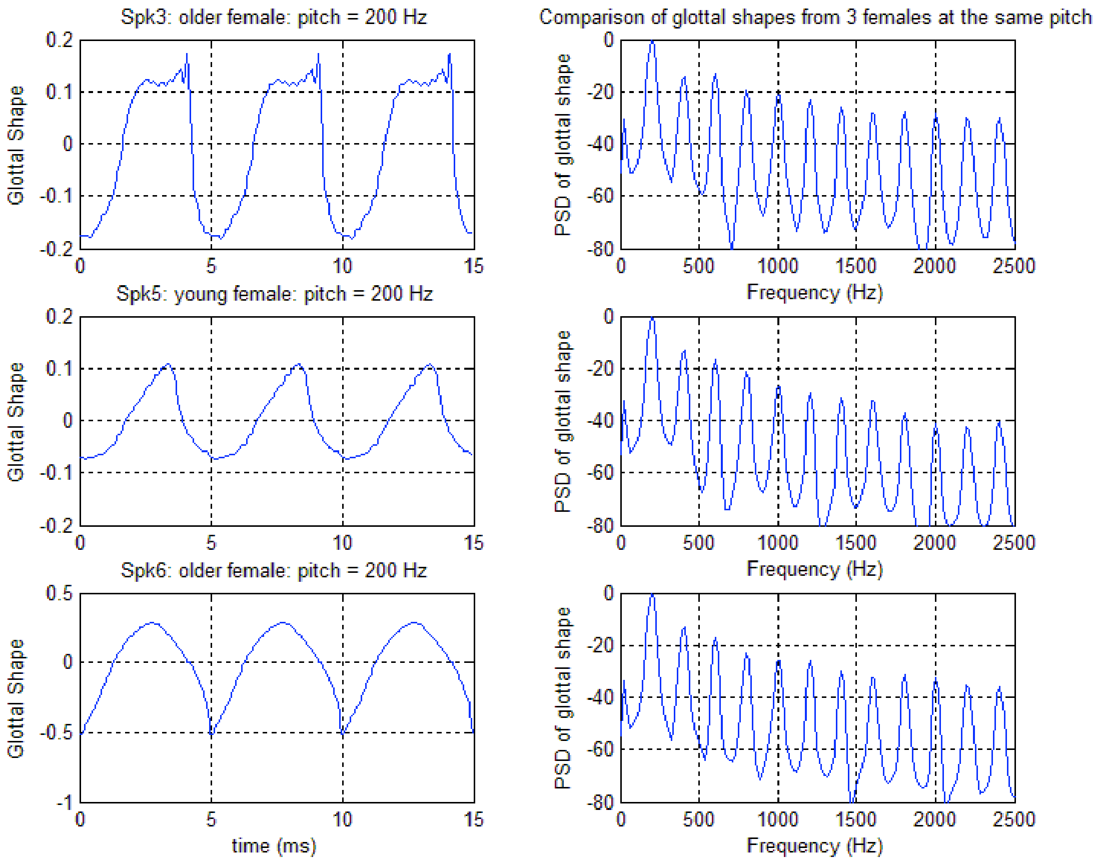


Fig T3-1: Glottal shape functions and PSDs from 3 female speakers (location 5 data) . Note the similarity of the PSD spectra.

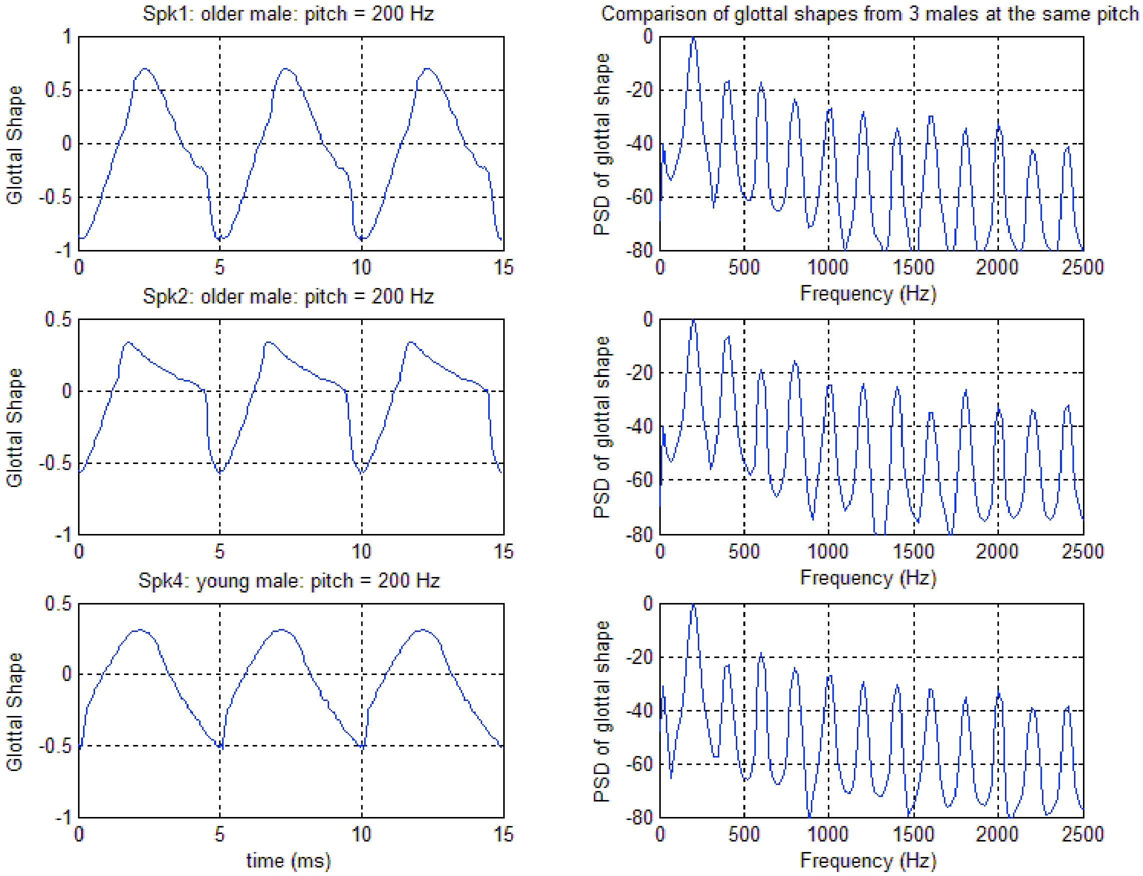


Fig. T3-2: Glottal shape functions from male speakers (position 5). Again notice the similarity of the spectral information, in spite of quite different real time shapes.

The above glottal data shown in T3-1 and T3-2 are often used directly to simulate air flow excitation functions because most of the important spectral information is contained in the “closure part” of the signal. That this a reasonable approximation is because the glottal signals are proportional to the area versus time of the closing glottis, which in turn is proportional to air flow (Flanagan 1965, Stevens 2000, Holzrichter et al 2003, 2005). For improved excitation functions, the radar response function must be removed (Burnett 1999), although modern EM sensors have quite wide bandwidths and have little signal distortion. In addition, EM sensor signals from surrounding tissues must be removed, which is a separate developmental issue. Nevertheless, for a great deal of survey work, of the type presented herein, the above approximation has been used.

Another glottal excitation approximation that is often used, because it can usually be easily simulated by delta-like functions, is a pressure excitation function (Burnett 1999). This is, to a first order approximation, the time derivative of the air flow excitation. See Fig. T3- 3 below, which is the pressure excitation corresponding to female speaker air flow excitation functions shown in Fig. T3-1 above.

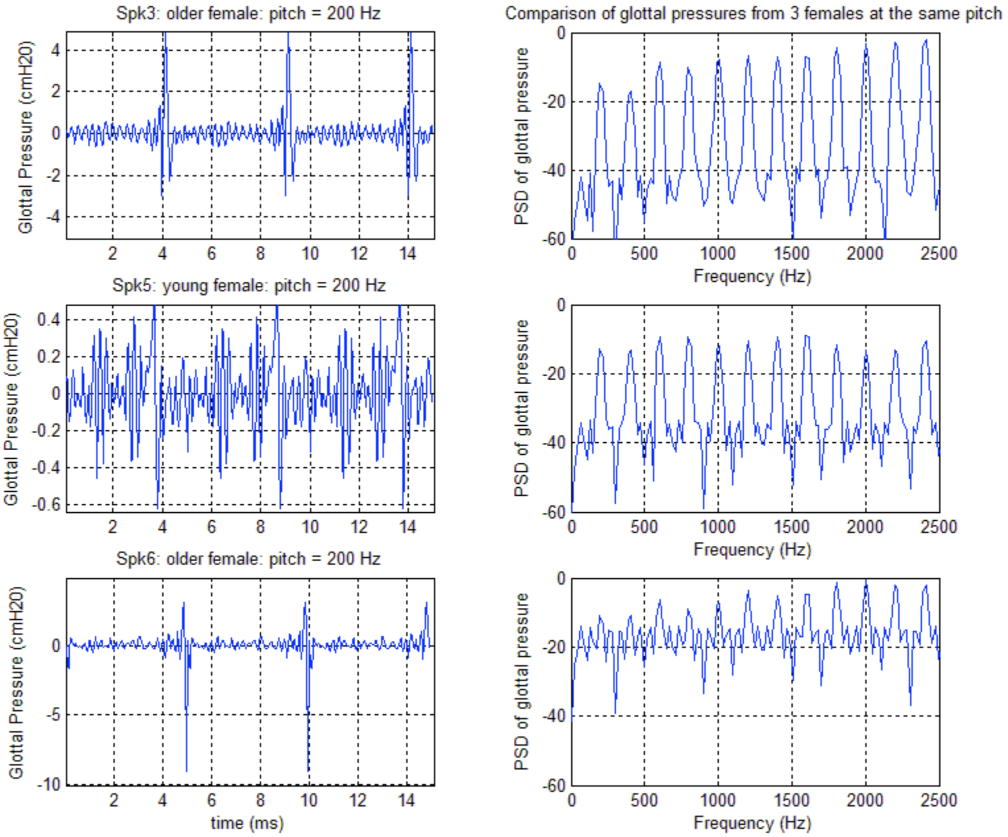


Fig. T3-3: Pressure excitation functions corresponding to the 3 air flow excitations from 3 female speakers shown above in T3-1. In principle they should be useful for inverting pressure induced tissue motion, but in practice acoustic resonator reflections and quasi-static pressure effects make this approach unusable now.

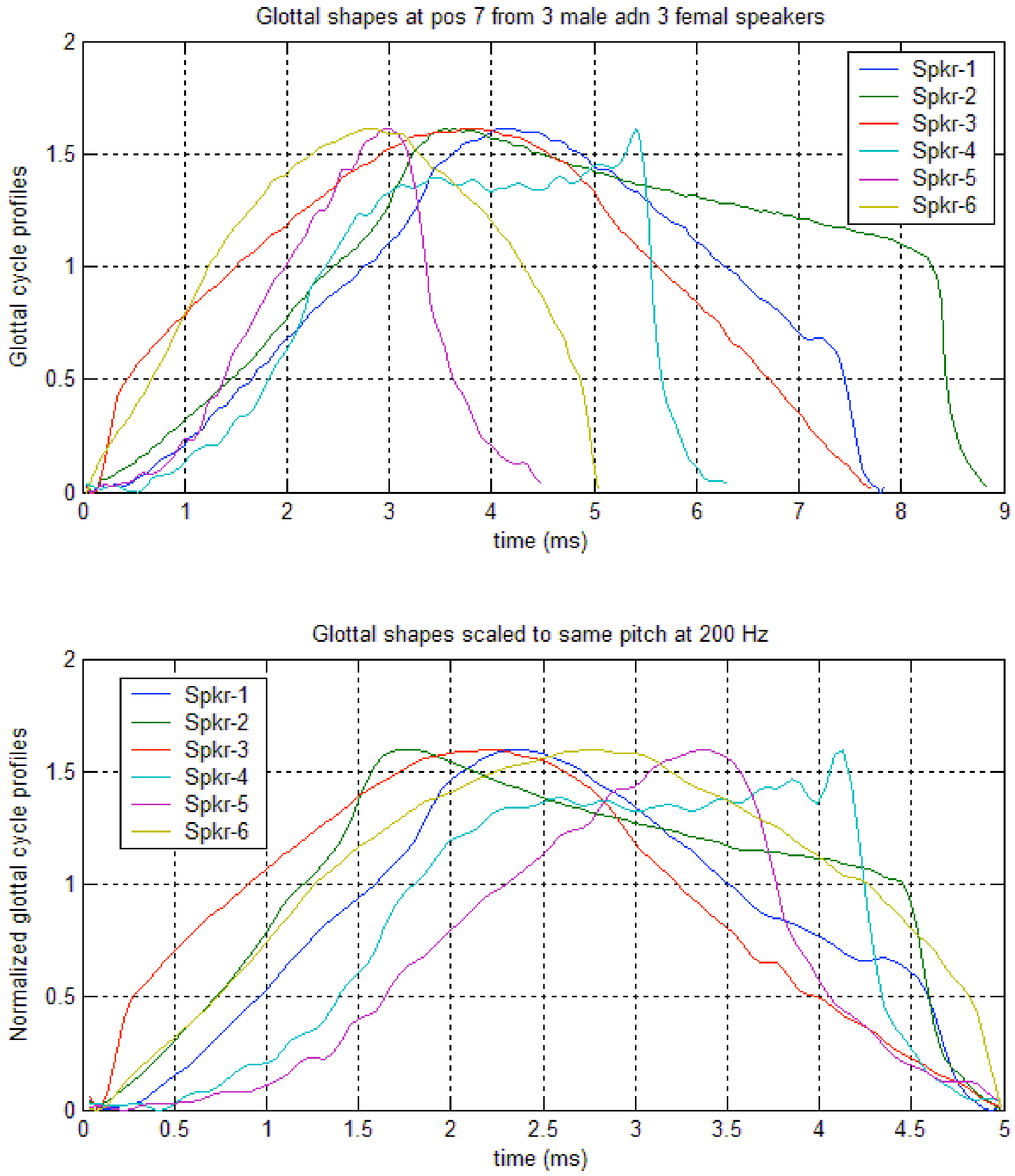


Fig. T3-4: Glottal shape summary of data from above figures T3-1 & 2, and the normalization of glottal shapes to a normalized pitch, (while preserving closure times) for storage in a catalogue of pitch shapes.

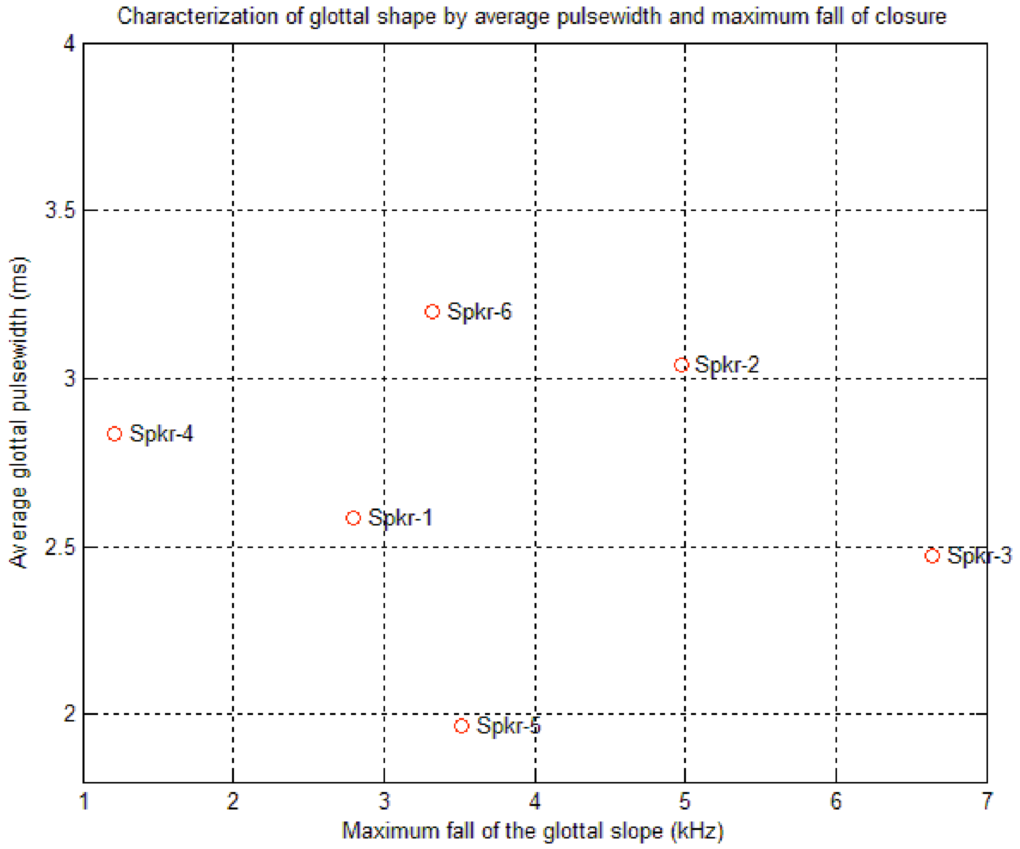


Fig. T3-5: Parameterization of glottal shapes by pitch and closure time. These indicate 2 of the parameters that were developed for glottal characterization.

Conclusion to Task 3:

The contractually proposed measurements and analyses for Task 3 have also been accomplished. These deliverables are improved data using a high sensitivity EM sensor and several antenna configurations, data from a representative number of speakers (8 used), comparison of glottal data to skin and tracheal wall data, and improved modeling of the excitation as volume air flow and also air pressure. Representative data and analyses meeting the commitments are shown primarily in the sections of this report labeled Sections I & III Introduction & Experimental configuration, Section IV EM Sensor and Antenna, and Section VI-3: Glottal Excitation Information.

Glottal region signals, from the vocal tract opening and closing, remain the best source of voiced excitation information. Additional signal vs location, polarization, bi-static sensing, amplitude levels and other data continue to confirm the source of signals from the glottal location at position 5 as being vocal fold opening and closing. Several very useful new characterization methods are described below.

VI-4: Task-4: Estimate sufficient excitation function:

The data and methods described in this report show a path to using available excitation information in the best possible way. In addition, there is a wide variety of data than can be obtained and used, when sensors, antennas, and algorithms are properly optimized.

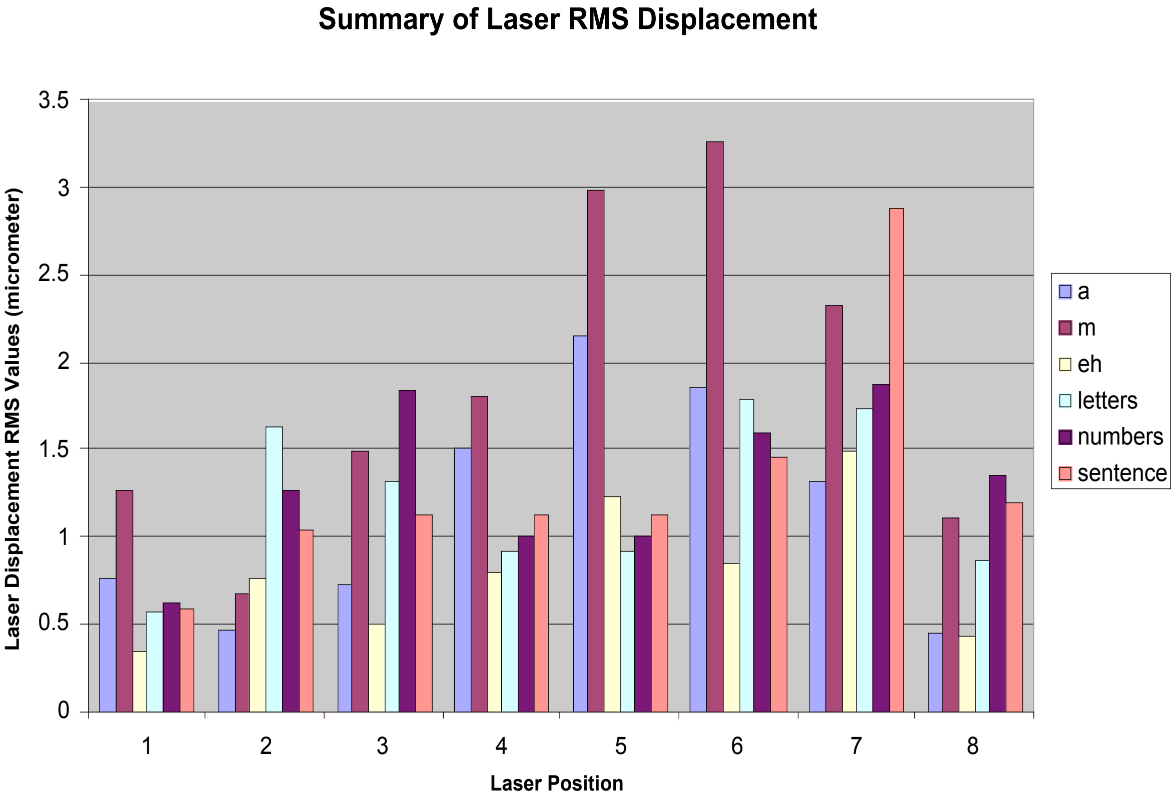


Fig. T4-1: Summary of skin surface displacements versus measuring location. On average it appears that location 7 (subglottal) has the most robust vibrations. As discussed in sections T1 and T2 above in the report, its tissues have the best spectral content for applications. Position 5 has moderate surface skin amplitude versus time, but very good internal tissue information (i.e., vocal fold motion information).

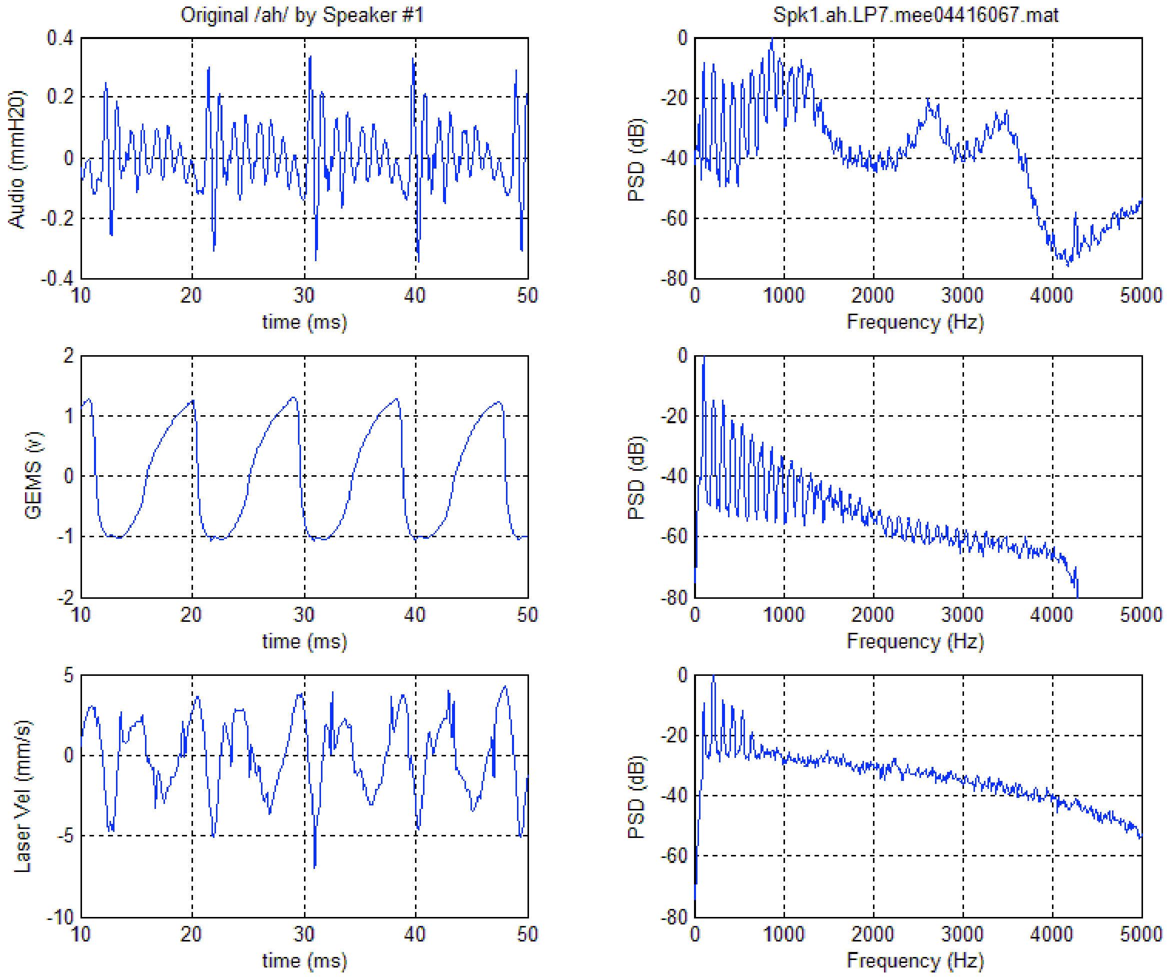


Fig. T4- 2 Example of a measured segment of speech for the voiced sound /ah/ , and its associated acoustic, glottal, and laser skin velocity data and PSD information. The noise in the velocity data is laser speckle noise from the relatively low level signal. These will be used to illustrate a newly developed method of defining a voiced excitation function, see below.

A numerical experiment was conducted using recorded data as shown just above. The intention was to see to what extent the limited data provided by laser detected skin motion from the subglottis (location 7) could be used to reconstruct a useable excitation function for denoising the acoustic signal.

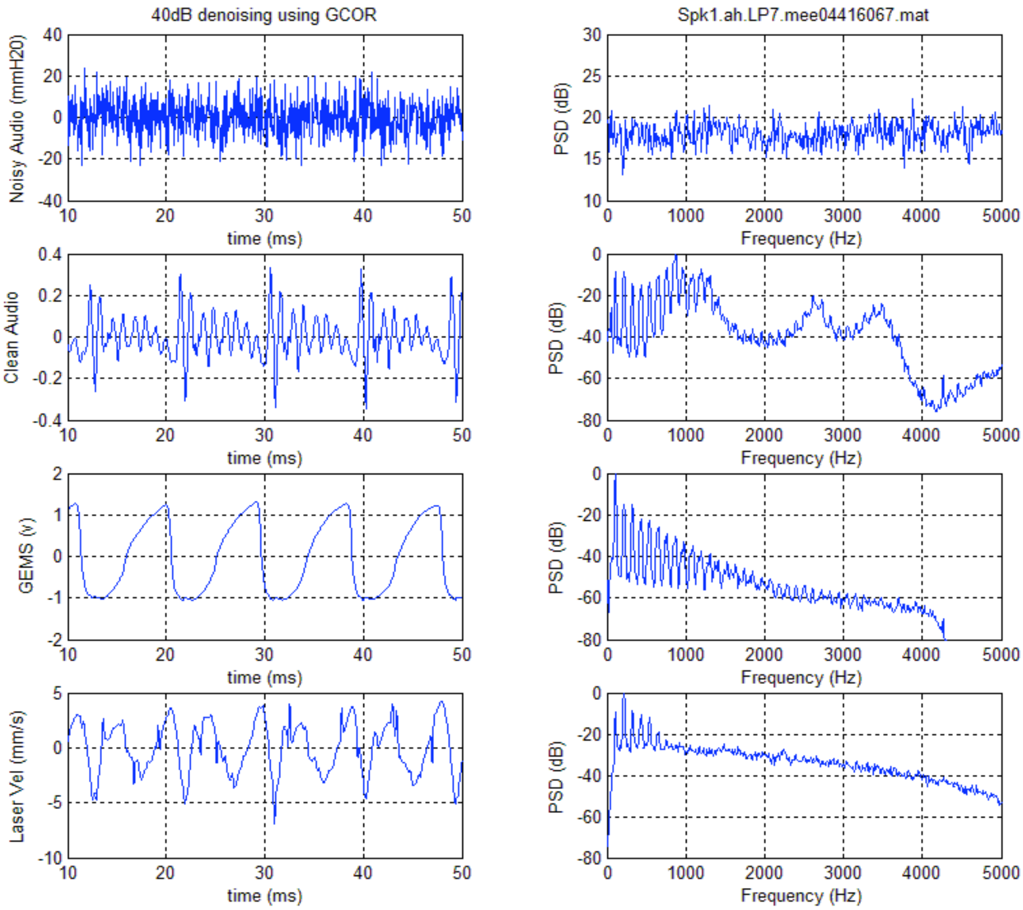


Fig. T4-3: Example of data shown in Fig. T4-2 (just above) but with 40 dB of white noised added to the acoustic signals. The second trace shows how the original glottal signal can be used to clean away the acoustic noise from the original voiced signal, ah , using the GCOR-EM sensor denoising technique invented by Ng et al (2000)

The GCOR approach (i.e., Glottal Correlation), using EM sensed glottal information and a background sensing acoustic microphone, was developed by Ng et al (2000) to use the frequency content of the excitation function to build a real time, spectral domain comb filter. This approach works very well in the case of white noise and also helicopter noise, see below. As discussed elsewhere (Ng et al 2000), other spectral approaches work well for coherent noise (out of band), and for “cocktail party” noise from nearby speakers with different pitch periods, and for situations where a noise sensing microphone is impractical. Other SNH developed techniques have been developed for use for cases of “in band” noise removal.

The illustration shown in Fig. T4-4, just below, shows how a prior record glottal function from the same individual’s information taken at a more recent time is used to accomplish GCOR denoising. Note the differently shaped glottal function, which has sufficiently similar spectral content to enable good noise removal. Trace 2 below, left and right sides, show the cleaned (denoised) signal from this individual.

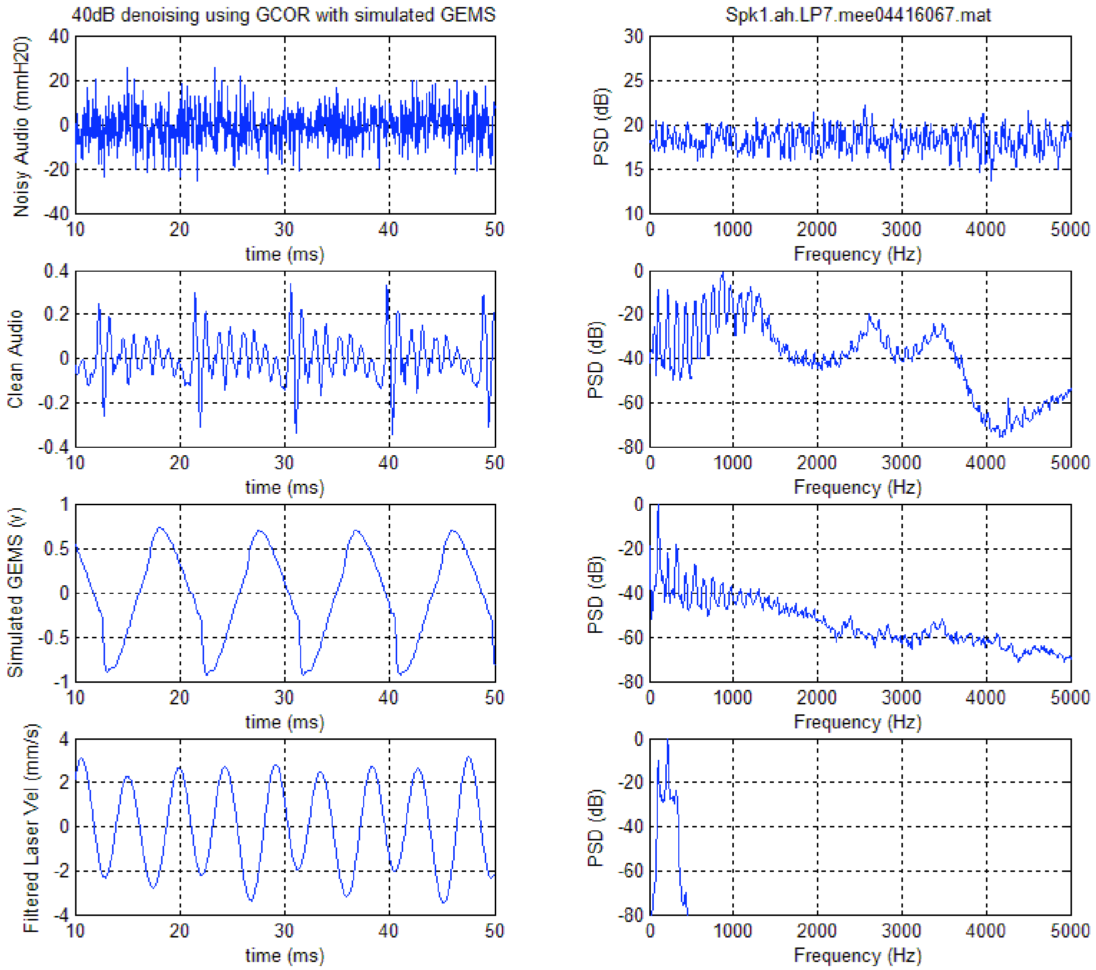


Fig. T4-4: Example of denoising as shown in the previous figure, only using the limited information in trace 4 to modify the stored glottal signal shown in trace 3.

In this example of denoising, the approach (see Fig. T4-5 below) uses a library of predefined glottal shapes. It was first necessary to find a previously recorded shape for this speaker’s excitation (or one sufficiently close to this speaker), then the stored shape was stretch to be the correct pitch, but retaining the proper closure rate, using the laser skin surface harmonics (shown on the right side of trace four in the above Fig T4-4. This function, trace 3 left side, was then used to perform the GCOR operations to remove the added 40 db of acoustic noise, see trace 2 right side. This example above used skin motion data, but tracheal tube motion data can be used as well. EM sensed tracheal data are noise immune and contain more harmonics, especially from the subglottal region, than skin data does.

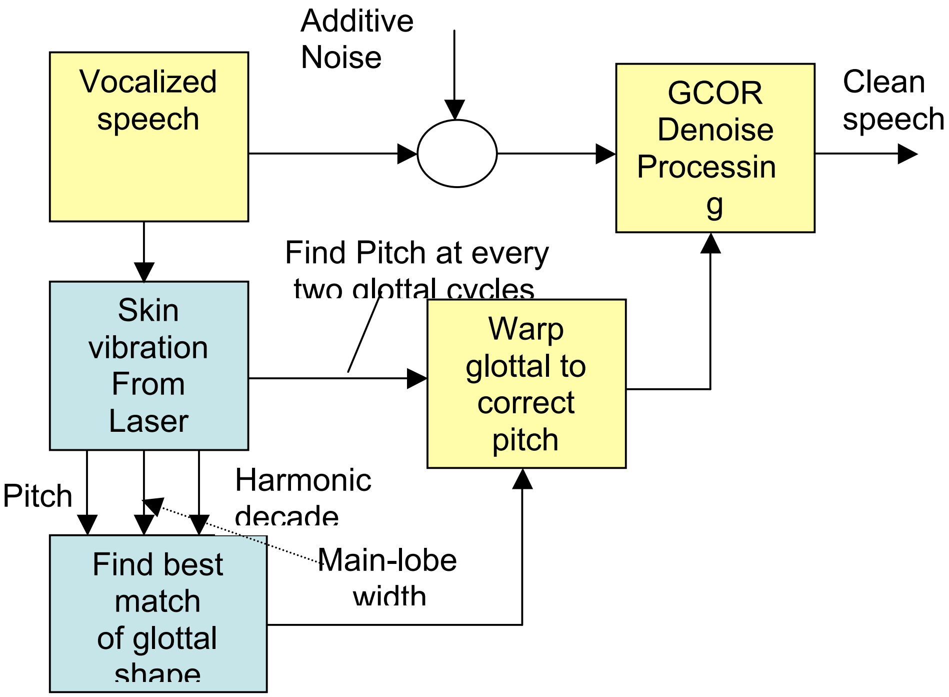

Additive
Noise
Vocalized speech
GCOR Denoise Processin g
Clean speech
Find Pitch at every two glottal cycles
Skin vibration From Laser
Warp glottal to correct pitch
Pitch
Harmonic decade
Main-lobe width
Find best match of glottal shape

Fig. T4-5: Algorithmic procedure for using spectral estimation, derived from limited vocal tract tissue and skin information, to determine a voiced excitation function. Denoising, illustrated in Figs T4-4, is one example chosen to demonstrate the effectiveness of these excitation functions.

Another test is the degree to which an acceptable transfer function can be obtained using a simulated glottal function. Fig. T4-7 below shows a simulation and ARMA fit to the transfer function, using denoised audio for ah and the simulated excitation function from Fig. T4-4, trace 3. The approximated data is converted to speech on an available CD, which is very understandable. However, more work needs to be done to develop approximations that carry needed speaker personality. This is produced by using the simulated excitation function twice, once for denosing and then for formant generation.

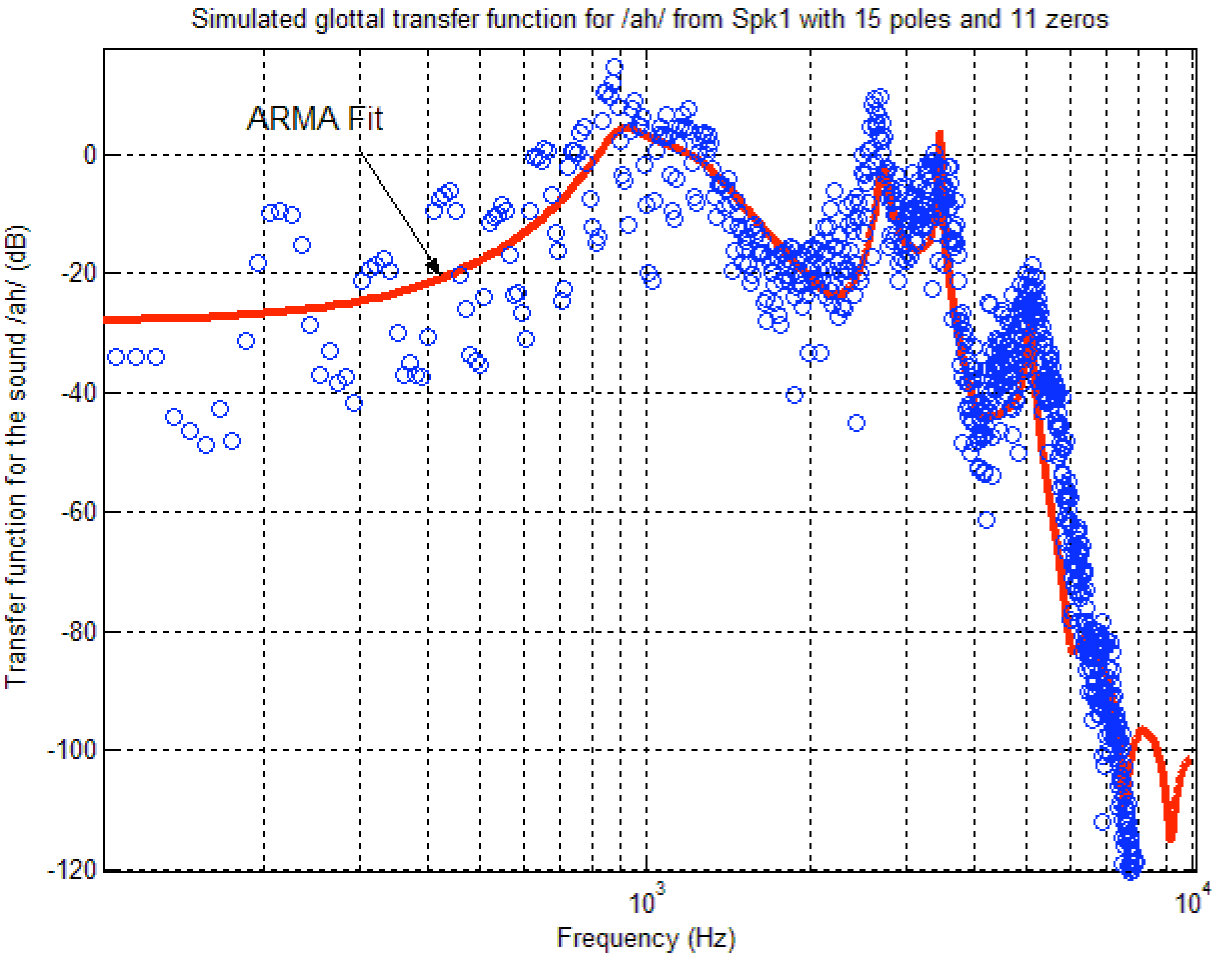


Fig. T4-6: Small circles show transfer function (point by point) for sound /ah/ using simulated glottal function and denoised audio sound /ah/ from Fig. T4-4. The solid line is a 15 pole, 11 zero fit to the data. The sound regenerated using the ARMA approximation is very understandable, but lacks some personality details.

Conclusion to Task 4:

The contractually proposed summaries and analyses for Task 4 have also been accomplished. These deliverables are to summarize the data, examine several voiced excitation function approaches, and give examples of how an estimated excitation function can be used for denoising and for voiced-speech formant definition. These are illustrated in the Task Sections T-1 to T-4. Representative data and analyses meeting the commitments for summarization are shown primarily in the sections of this report labeled Section I Introduction, several summaries of data are illustrated in the sections VI-1 to V1-3, and models and examples are discussed in Section VI-4: Estimates of glottal Excitation Function. A CD is available which has the raw data used in this report.

EM sensor obtained glottal closure data is the most robust source of voiced-excitation time and spectral information for Darpa’ s and many other applications. These data are somewhat more difficult to obtain in a real-life environment because of the necessary for careful positioning of an EM sensor at a point on the neck that does not easily support a small sensor. Hence other sources of excitation information have been investigated for this report—namely skin surface vibrations and tracheal tube vibrations. We have found that these other sources of information can also be used to generate quite a good voiced excitation function, which as shown above, can accomplish real time incoherent noise removal from the speaker’s acoustic signal. Skin surface vibrations can be measured at many locations with limited frequency response, with the subglottal, on-axis location (#7) providing the richest and best information. It is important to keep in mind that the skin vibrations are susceptible to loud external noise. Tracheal tube vibrations, measured with EM sensors, inside the neck provide improved information compared to skin data, but not as much as direct glottal excitation data. The superglottal tissues of the trachea are limited in their spectral content, while those at the subglottal region have more spectral content.

Denoising and formant characterization have been demonstrated with the three types of data – skin, vocal tract walls, and vocal folds. All of these data can be used with good success to meet the needs of the Darpa Vocoding in High Noise programs, as well as other speech characterization applications.

VII: Summary

This contract has enabled a good general understanding of neck and cheek skin response to external and internal acoustic pressure pulsations. In particular we have investigated the capacity of skin, trachea tube, and vocal folds to convey spectral data of sufficient usefulness for Darpa's , and others', speech characterization applications. We have investigated the imitations of inverse transforms from skin to excitation functions, and the spectral information in tissues themselves. We have also discovered that there is a way to estimate sufficiently good voiced-excitation functions using the data that is available from tissue vibration data whose spectral data is limited to below I kHz (in some cases below 500 Hz). In addition we have continued to obtain information on EM sensed vocal fold motions, which provide full-spectrum excitation functions. The report sections on the specific tasks 1-4, contain details of the summarized conclusions in this conclusion section.

We quantified the impact of loud external noise on the vibration properties of skin (notable influence) and internal tissues (no influence), and on several different types of sensors—EM, laser, and accelerometer. Noise does not affect these sensors per se, but it can cause spurious skin vibrations. These can confuse sensors relying on low level vibrations, such as accelerometers, microphones, strain gauges, etc.

We have been able to compare 3 different types of EM sensors and how they measure amplitude, phase, spectrum, gain, etc. This is very useful as it tells us how various types of sensors can be developed for various applications. In addition we investigated three different antenna types –waveguide, dipole, and patch- transmitting and receiving in several different polarizations, as well as investigating monostatic (one antenna for both transmit and receive) and bistatic conditions (two antennas for transmit and receiving with varying relative polarizations). In addition, we believe that we can now optimize sensor and antenna designs for almost all anticipated fielding configurations—ranging from chin strap mounting, to necklace mounting, to neck armor mounting, and others.

Finally, we developed a methodology for estimating voiced excitation functions with sufficient spectral information to enable narrow bandwidth vocoding and to enable the use of spectral noise removal techniques. From denoised audio signals we are then able to estimate good formants (using ARMA techniques) for vocoding and other applications. We have shown that excitation information from the subglottal region is especially robust, and has sufficient information to define a good voiced excitation function. We feel that these methods of extracting every bit of information from limited excitation information are very exciting and will be very valuable.

VIII: Future Work:

Two general topics for future investigation are worth considering, and would have large, long term payoffs. They are:

1) Obtain information and demonstrate the concepts of extracting all needed voiced excitation function using the first 2 to 3 harmonics from tissue vibrations. In addition, explore the range of types of excitation shapes for catalog look-up, and determine the best methods for matching measured (real time) harmonic spectral information to the appropriate catalogued excitation shape. Determine the extent to which sub-glottal and superglottal tissue vibration information can be used to define the voiced excitation.

2) Develop and demonstrate concepts for specialized EM sensor antennas to extract as much excitation information as possible from vocal fold and trachea tube excitations. In particular, develop an antenna that relies on skin conduction of EM waves from an antenna element, located at some distance from the glottal region, to obtain more complete spectral information from the vocal folds.

3) Demonstrate new vocoding and denoising concepts, developed by SNH personnel, using the newly developed excitation function techniques, based upon estimating spectral information, as described in this report. Do so for several locations specified by Darpa personnel.

VIV: Acknowledgements:

We would like to thank Prof. Sunil Puria of Stanford University for his assistance on many of the measurements in this report and Dr. Edward Damrose of the Stanford University Medical Center for his observations from a clinical perspective. We thank Mr. Fred Beckner and Mr. Steve Davis of Cyberdyamics Corp. in Palo Alto for radar and antenna work. We also wish to thank Drs. John Rosowski of the Peabody Laboratory at Mass. Eye and Ear Infirmary and Drs. James Kobler, at Mass General Hospital, for their many years of assistance. We would also like to thank Dr. Harold Cheyene of Mass Eye and Ear and Sensimetrics Inc for his assistance.

Bibliography:

Burnett, G.C.,(1999) “The Physiological Basis of Glottal Electromagnetic Micropower Sensors (GEMS) and Their Use in Defining an Excitation Function for the Human Vocal Tract” Thesis UC Davis, Jan. 15th, 1999, available through ProQuest Digital Dissertations, document number 9925723.

Cheyne 2002 “Estimateing glottal voicing source characteristics by measuring and modeling the acceleration of the skin on the neck” H.A.Cheyne II, Thesis MIT, February 2002.

Flanagan, J.L. 1965 “Speech Analysis Synthesis and Perception”
Springer/Academic Press NY, Berlin.

Fung, Y.C. (1993) "Biomechanics - Mechanical Properties of Living Tissues" , p. 474, 2nd ed. (1993) Springer, NY

Holzrichter, J.F. (1995) “New Ideas for Speech Recognition and Related Technologies”, Lawrence Livermore National Laboratory Report, UCRL-UR-120310 , 1995 . Available from Lawrence Livermore National Laboratory library, or from NTIS in Springfield, VA at http://www.ntis.gov/ordering.htm

Holzrichter, J.F., Burnett, G.C., Ng, L.C., and Lea,W.A. (1998) "Speech Articulator Measurements Using Low Power EM Wave Sensor" , J. Acoust. Soc. Am. **103** (1) 622, 1998. Also see the Website http://speech.llnl.gov/

Holzrichter J.F and Burnett, G.C. (1999) “Human Speech Articulator measurements using low power, 2 GHz Homodyne Sensors” 24th international conference on infrared and millimeter waves”, Lombardo. L. A. ed, Monterey, CA. Sept 5 (1999), Kluwer Publishing; see also Lawrence Livermore Laboratory report UCRL-JC-134775M

Holzrichter 2003, Holzrichter J.F., Kobler, J. B., Rosowski, J.J., Burke, G.J., (2003) “EM wave simulations and measurements of glottal structure dynamics – a Technical Report” UC Report UCRL-JC-147775 , 2003. Available on the Website http://speech.llnl.gov/ or from Lawrence Livermore National Laboratory library, or from NTIS in Springfield, VA at http://www.ntis.gov/ordering.htm

Holzrichter et al., (2005) , Holzrichter, J.F.; Ng, L.C.: Burke, G.J. Champagne N.J; Kallman, J.S; Sharpe R.M; Kobler, J.B; Hillman, R.E.; Rosowski , J.J “Measurement of Glottal Structure Dynamics” JASA , (Feb 2005) .

Ishizaka, K., French, J.C., and Flanagan, J.L., (1975) “Direct Determination of Vocal Tract Wall Impedance,” IEEE Trans. Acoustics, Speech, and Signal Processing **ASSP-11** (4), 370 (1975)

McEwan, T.E., (1994) U.S. Patents 5,345,471, 5,361,070, and 5,573,012

Meltzner, G.S.; Kobler, J.B; Hillman, R.E. (2003) “Measuring the neck frequency response function of laryngectomy patients: Implications for the design of electrolarynx devices” JASA 114 (2), 1035 (August 2003).

Ng, L. C.; Burnett, G. C.; Holzrichter, J. F.; and Gable, T. J. (2000) “De-noising of Human Speech Using Combined Acoustic and EM Sensor Signal Processing”, Icassp-2000, Istanbul, Turkey, June 6, 2000 . Available at http://speech.llnl.gov

Ng, L. C.; Holzrichter, J. F., and Larsen, P.; (2002) “Low Bandwidth Vocoding using EM Sensor and Acoustic Signal Processing”, submitted to Icassp-2001, Available at http://speech.llnl.gov

Skolnik, M. (1990). “*Radar Handbook*,” 2nd edition., McGraw-Hill, New York

Stevens, K.N. (2000) “Acoustic Phonetics” MIT Press, Cambridge, MA

Titze, I.R., (1994)“Principles of Voice Production” Prentice Hall, NJ, 1994

Titze, I.R., Story, B.H., Burnett, G.C., Holzrichter, J.F., Ng, L.C., Lea, W.A., (2000) “Comparison between electroglottography and electromagnetic glottography” J. Acoust. Soc. Am. **107** (1), 581 (2000)