Preprint
UCRL-JC-140496

Speech Articulator and User Gesture Measurements Using Micropower, Interferometric EM-Sensors

J.F. Holzrichter

This article was submitted to
2001 Institute of Electrical and Electronics Engineers
Instrumentation and Measurement Technology Conference,
Budapest, Hungary, May 21-23, 2001

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

September 15, 2000

Speech Articulator and User Gesture Measurements Using Micropower, Interferometric EM-Sensors

J.F. Holzrichter
Lawrence Livermore National Laboratory and University of California, Davis
holzrichter1@llnl.gov

Key words: radar, microwaves, speech, coding, microphones.
Preferred Presentation: Oral

Abstract

Very low power, GHz frequency, "radar-like" sensors can measure a variety of motions produced by a human user of machine interface devices. These data can be obtained "at a distance" and can measure "hidden" structures. Measurements range from acoustic induced, 10-micron amplitude vibrations of vocal tract tissues, to few centimeter human speech articulator motions, to meter-class motions of the head, hands, or entire body. These EM sensors measure "fringe motions" as reflected EM waves are mixed with a local (homodyne) reference wave. These data, when processed using models of the system being measured, provide real time states of interface positions vs. time. An example is speech articulator positions vs. time in the user's body. This information appears to be useful for a surprisingly wide range of applications ranging from speech coding and recognition, speaker or object identification, noise cancellation, hand or head motions for cursor direction, and other applications.

Introduction

Recent studies using micro power radar-like sensors have shown that human (i.e., animate) speech articulator motions (see Holzrichter et al. [1]) and inanimate mechanical vibrations can be measured in real time as their corresponding acoustic sounds, such as speech or musical instrument sounds, are produced. Initial work also showed that very simple, non-spatially localized measurements of speech articulators can provide information on a wide variety of motions associated with speech sound production—such as tissues associated with the glottal region, jaw, tongue, soft palate, lips and others. Similarly, characteristics of vibrating mechanical structures such as musical instrument strings, and vibrating plates are easily measured. Larger motions of a use's body parts are also being measured using "multiple fringe" counting, together with fractional fringe techniques appropriate for homodyne radar-like sensors.

There are three primary modes of detection: 1) small scale movements where Δx is $<< \lambda$; 2) intermediate scale where $\Delta x < \lambda/4$; and 3) where $\Delta x > \lambda/4$. Because the electronics in these sensors do not work as well at low frequency (e.g., DC) as they do at higher frequencies (e.g., > 20 Hz), and because of the "cluttered" environment in which these applications commonly take place (i.e., inside the human head), applications have concentrated on those structures that reflect EM waves at rates significantly greater than 5 Hz. Experiments to date have concentrated on measuring specially "targeted" structures such as

the human glottal system, the jaw/tongue system, and specially modulated reflectors. Modulated microwave reflectors are used for calibration or for attachment to a body part to enable specifically targeted local motion measurement.

Homodyne Sensors

The homodyne field disturbance EM sensor works as an EM wave interferometer by comparing the reflection of a transmitted wave against a local (phase reference) wave. As the targeted reflecting surface of the system moves, the phase of the reflected wave varies with respect to the stationary local wave. A signal associated with the product of the two waves is detected by a mixer, is integrated, amplified, and filtered (see Fig. 2). The detection system and internal filters measure position change (as small as 2 micrometers) versus time, within specific frequency bands ranging from <5 Hz to 7 kHz. The homodyne response function for a GEMs sensor designed to sense glottal tissue motions at 70-7kHz, is shown in Fig. 3 (GEMs stands for glottal EM sensor). The sensor response (see Fig. 3) was measured using a 100 cm^2 Cu plate, vibrating at 200 Hz with an amplitude of about 100 micrometers, that was moved to a new location for each data point, see Burnett [3]. When these data are used with non-unity dielectric systems, (e.g., human tissue with $\varepsilon \approx 50$) the corresponding effective wave path must be taken into account. The thin arrow in Fig. 2, at 70 mm, shows the effective position on the sensitivity curve of the rear windpipe wall of a subject. The large number arrows illustrate three different modes of using these sensors: 1) small distance, 2) medium distance, 3) large distance.

Excitation and Transfer Functions

A particularly interesting application of these sensors is to the measurement of wind pipe wall motion, i.e., amplitude versus time, in response to the subglottal pressure changes as the vocal folds open and close. See Fig. 3. The association of the EM sensor signal with vibrations of these particular tissue-air interfaces, was accomplished by moving the sensor relative to the neck (see Fig. 3) and noting the signal nulls and signal sign changes. These data enable an estimation of the voiced pressure excitation function of human speech to be made [3]. For each excitation signal point, there is corresponding acoustic signal that reaches a standard acoustic microphone at a delayed time (about 1.3 msec). The excitation signal generated at the glottis is reflected and absorbed as it travels up the vocal tract and out to the microphone. The reflecting and transmitting properties of the vocal tract are a function of the tract's length and internal structures (i.e., its resonant and absorbing cavities). Using well-established signal processing techniques, the filtering effects of the tract, between the glottal source and the microphone can be estimated when the excitation function is known. The data insert in Fig. 4 shows an example of the corresponding transfer function (TF) power spectrum taken as the sound /a/ (pronounced "ahh") was spoken. These methods enable more accurate "pole/zero" approximations of the TF to be used, than presently used LPC methods. From the shape of the power spectrum of the transfer function, speech recognition algorithms can identify the sound being spoken as the phoneme /a/. Conversely, once excitation and transfer function information for an individually pronounced sound element is obtained, the sound can be reconstructed, i.e., synthesized. Similar experiments have been performed on stringed instruments, mechanical structures, and many other human speech sounds that have "source-filter"

characteristics. The additional data obtained with excitation-measuring radar-like sensors can be used to more accurately characterize the sounds being produced, to synthesize their sounds, to cancel sounds, or to filter background noise.

Human Use

The total radiated output of presently used EM sensors is 0.3 milliwatts, into 4 pi sterradians. When the sensor is placed close to the skin, about 1/2 of the output reaches the skin over an area of about 1.5 cm^2. Thus the average power on the tissue is < 0.1 mwatt/cm^2. This level is well below the U.S. standard for continuous user exposure to EM waves of 1.0 mW/cm^2 and is consistent with Swedish and Finnish standards of 0.1 mW/cm^2 for continuous exposure. In comparison, cellular telephones radiate about 1 W in similar frequency bands similar to those used with these EM sensors, and thus about 1 to 10 mW/cm^2 reach the user's head and facial region, intermittently. The homodyne EM sensors described herein can be built into generally used appliances and communication devices, without any adverse effects to the user or to bystanders.

Conclusion

Low power EM radar-like sensors can measure generalized motions of animate and inanimate objects accurately, safely, and compactly. They are especially useful for measuring motion parameters of system components that are obscured by other dielectric materials, that are extended in dimension, when rapid (i.e., speech-of-light) measurement is needed. For applications near human users, their low power output is very desirable. In particular, the homodyne EM sensor is proving to be very useful for measuring air-tissue interface motions of system excitation sources, as sound is produced. These enable simple and detailed descriptions of the acoustic behavior of human speech systems and a wide variety of sounds producing mechanical systems. In addition, their large-scale motion measurement properties are being applied to more conventional user interface devices such as "pointer" and "mouse-like" devices.

Acknowledgements

The authors would like to thank Drs. L. C. Ng and G. C. Burnett, and E. T. Rosenbury and T. A. Gable for their assistance. Advice from Professors N. Luhmann and R. Freeman are appreciated. The author thanks the U.S. Department of Energy and the National Science Foundation for their support. This work was performed in part under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

References

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

[2] McEwan, T.E., U.S. Patent No. 5,345,471 (1994), U.S. Patent No. 5,361,070 (1994). U.S. Patent No. 5,573,012 (1996).

[3] Burnett, G.C., "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 on the Website mentioned in [1], and through University Microfilms, Ann Arbor, MI, document number 9925723.

Figures

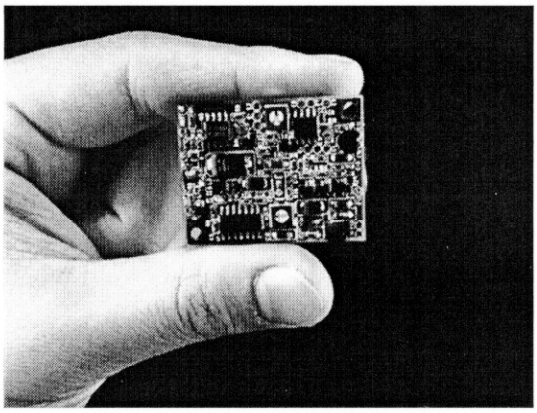

Figure 1. This microwave sensor transmits nominal 10 cycle pulse trains at frequencies ranging from 900 MHz to 5 GHz. It uses a homodyne receiver mode, radiates < 0.3 mwatts of power, and uses internal filtering to detect motions cycling at > 10 Hz. Two patch antennas, 1.5 cm x 1 cm, are used to transmit and to receive the EM waves. With a high-gain antenna, such devices operating at 4 GHz have a > 50M range.

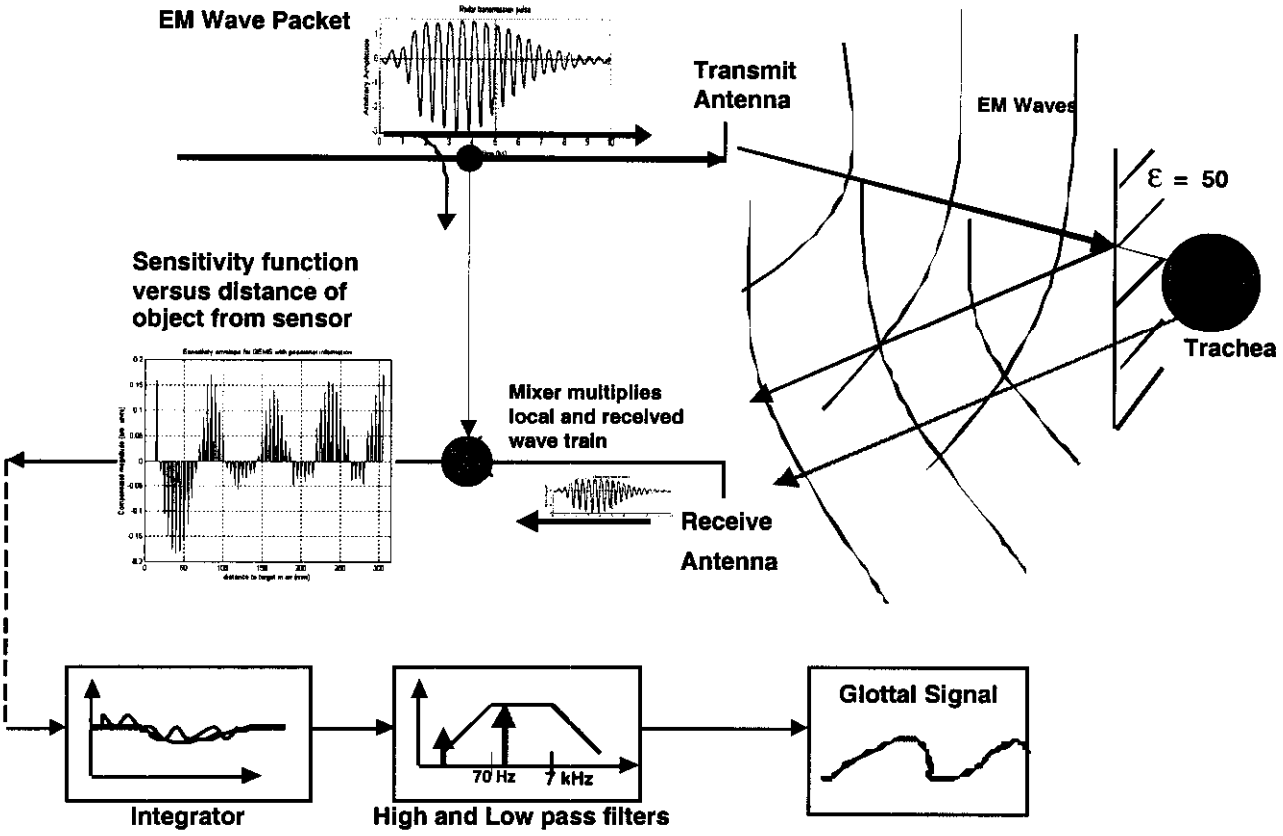

Figure 2. Typical homodyne circuit when used in a smell-signal, field-disturbance mode to measure glottal tissue motions. A wave packet is generated by a gated oscillator (upper left) and is both transmitted to a target (e.g., human trachea) and directed to a diode mixer where it is "multiplied" times the signal that is reflected and received by a second antenna. The received signal's low-frequency envelope (from the mixer) is integrated, filtered, and then processed to described interface distance versus time.

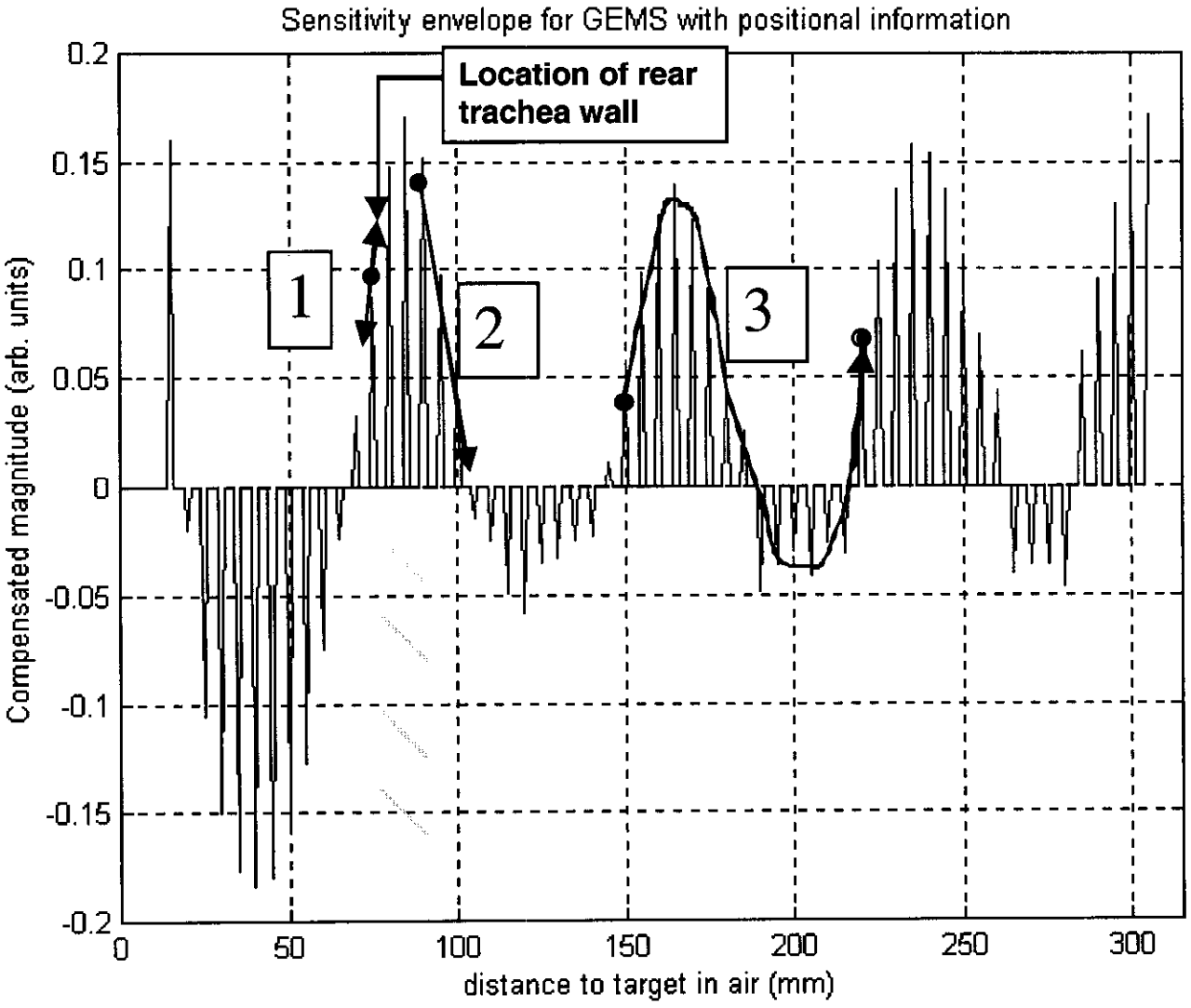

Figure 3. Homodyne sensor sensitivity curve as measured using a vibrating (~ 100 μm amplitude at 200 Hz) metal plate versus distance. Location 1 indicates the sensor's small signal response of the rear tracheal wall motion (~ 50 μm) of the user. From this data, speech excitation pressure is obtained. Location 2 indicates larger scale motion of a speech articulator, such as the 1 cm to 2 cm jaw-to-palate distance. Location 3 illustrates how > 7 cm motion of a user-interface device would be "tracked" and converted to relative distance traveled.

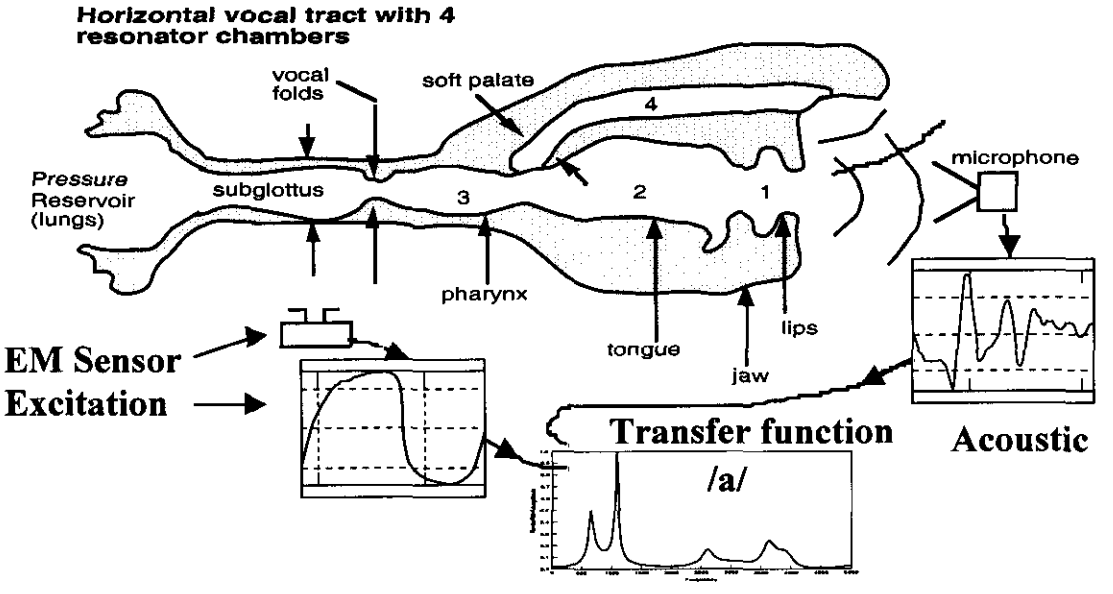

Figure 4. Midsagital plane of "linearized" human vocal system showing speech articulators, and EM sensor measurement locations. The GEMs EM-sensor measures tissue motions as a consequence of subglottal pressure rise and fall, as the vocal folds open and close. Signals include the excitation function, acoustic speech signal, and the transfer function (for the sound /a/) derived by deconvolution of the excitation from the output.