

# **New Ideas for Speech Recognition and Related Technologies**

*J. F. Holzrichter*

**June 17, 2002**

*U.S. Department of Energy*

Lawrence  
Livermore  
National  
Laboratory

## DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed 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.

This report has been reproduced directly from the best available copy.

Available electronically at <http://www.doc.gov/bridge>

Available for a processing fee to U.S. Department of Energy  
And its contractors in paper from  
U.S. Department of Energy  
Office of Scientific and Technical Information  
P.O. Box 62  
Oak Ridge, TN 37831-0062  
Telephone: (865) 576-8401  
Facsimile: (865) 576-5728  
E-mail: [reports@adonis.osti.gov](mailto:reports@adonis.osti.gov)

Available for the sale to the public from  
U.S. Department of Commerce  
National Technical Information Service  
5285 Port Royal Road  
Springfield, VA 22161  
Telephone: (800) 553-6847  
Facsimile: (703) 605-6900  
E-mail: [orders@ntis.fedworld.gov](mailto:orders@ntis.fedworld.gov)  
Online ordering: <http://www.ntis.gov/ordering.htm>

OR

Lawrence Livermore National Laboratory  
Technical Information Department's Digital Library  
<http://www.llnl.gov/tid/Library.html>

## FOREWORD

The ideas relating to the use of organ motion sensors for the purposes of speech recognition were first described by the author in spring 1994. During the past year, a series of productive collaborations between the author, Tom McEwan and Larry Ng ensued and have lead to demonstrations, new sensor ideas, and algorithmic descriptions of a large number of speech recognition concepts. This document summarizes the basic concepts of recognizing speech once organ motions have been obtained. Micro power radars and their uses for the measurement of body organ motions, such as those of the heart and lungs, have been demonstrated by Tom McEwan over the past two years. McEwan and I conducted a series of experiments, using these instruments, on vocal organ motions beginning in late spring, during which we observed motions of vocal folds (i.e., cords), tongue, jaw, and related organs that are very useful for speech recognition and other purposes. These will be reviewed in a separate paper.

Since late summer 1994, Lawrence Ng and I have worked to make many of the initial recognition ideas more rigorous and to investigate the applications of these new ideas to new speech recognition algorithms, to speech coding, and to speech synthesis. I introduce some of those ideas in section IV of this document, and we describe them more completely in the document following this one, UCRL-UR-120311. For the design and operation of micro-power radars and their application to body organ motions, the reader may contact Tom McEwan directly.

The capability for using EM sensors (i.e., radar units) to measure body organ motions and positions has been available for decades. Impediments to their use appear to have been size, excessive power, lack of resolution, and lack of understanding of the value of organ motion measurements, especially as applied to speech related technologies. However, with the invention of very low power, portable systems as demonstrated by McEwan at LLNL researchers have begun to think differently about practical applications of such radars. In particular, his demonstrations of heart and lung motions have opened up many new areas of application for human and animal measurements.

The author acknowledges the very important contributions of Tom McEwan and Larry Ng in bringing these ideas to their present level so rapidly. In addition many important contributions have been made by my colleagues: Pat Welsh on electronics, Ursula Goldstein and Mike Portnoff on issues in speech recognition, Noel Sewall on data acquisition, and Dr. Wayne Lea of the Speech Science Institute for commenting on this document. Finally, I would like to thank the management of LLNL's Laser Program, Engineering Department, and the Director's Office for support during the initial stages of this work.

## TABLE OF CONTENTS

- I). Introduction to the topic of Speech Recognition augmented by EM Sensors
- II. EM Sensors and Signal Processing
- III. Single organ-motion algorithms in conjunction with acoustic speech
  - a) whole organ motion EM sensing
  - b) partial organ motion EM sensing
- IV. Speech tract model based algorithms
- V. Multiple organ - multiple position EM data algorithms:
  - a) with acoustic speech
  - b) without acoustic speech
- VI. Word signature algorithms
  - a) one sensor EM data
  - b) multi-sensor EM data
  - c) natural language recognition
- VII. Conclusions and Other Applications

# NEW IDEAS IN SPEECH RECOGNITION

JOHN F. HOLZRICHTER

## I. Introduction

In April 1994, the author described a method of non-acoustic speech recognition and shortly thereafter, the author and Tom McEwan described and validated electromagnetic techniques to obtain speech organ position and velocity information. For a timely, authoritative review of the status of speech recognition and other technologies, see reference 1, which is a publication of the National Academy of Sciences. The original descriptions identified simple algorithms for reducing the organ position information to a set of position vectors for each speech unit. By speech unit, I mean a word, a syllable, or a phoneme of sound that is defined to be the minimal unit to be recognized. See Appendix A for a list of American English phonemes given by Rabiner (2). The ideas of speech recognition using radar technologies is valuable and needed because conventional acoustic speech recognition of the English language is insufficiently accurate for most applications. Under the best circumstances one sees >5% error rates, and one typically sees >10% with natural language in adverse (e.g. noisy) environments. These make present all acoustic recognition systems not yet adequate for use in high value and/or in high noise conditions, and not acceptable for most users involved in work using large vocabularies. In addition, present recognition systems have difficulty with dialectal or foreign speakers. The purpose of this document is to review for the reader the basic ideas behind using electromagnetic wave scattering (i.e. radar scattering) from body organs to enhance speech recognition. I briefly review the principles of radar scattering from human organs, and then I describe a series of speech recognition algorithms that have been devised in order to make possible the applications of the above described ideas to a wide range of sensor and user conditions. In the conclusion, I speculate on other applications of this technology.

The important aspect of joining acoustic measurements and EM wave scattering measurements is that the two techniques are statistically independent, in a measurement sense, from one another. As a result the statistical accuracy of the non-acoustic techniques can be joined with the statistical accuracy of the conventional acoustic techniques to provide an improved method of accurately recognizing human speech. For example, there are many sounds that can be identified by EM scattering alone, which illustrates this notion of statistical independence (it also makes possible the concept of "soundless" speech recognition). There is another class of spoken words that are acoustically difficult to differentiate from one another (e.g. "saline" and "sailing"), but through EM scattering from the participating speech organ(s), are quite distinctly recognizable. This "saline etc." example is

differentiated through tongue-palate position measurements. See reference 3 by Olive et al for a review of speech organ sound relationships.

Our initial measurements indicate that recognition of needed organ positions or motions can be accomplished with high accuracy (exceeding 99% depending on the measurements). We also note that many of the errors contributing to the present 95% acoustic recognition accuracy can be readily corrected with the addition of EM sensor data, and thus the joint accuracy can be improved to better than 99%. This level begins to approach the quality of present human recognition of speech. In summary, the non-acoustic algorithms described in this document can be used separately to identify words (completely non acoustically) or can be joined with conventional acoustic recognition techniques to yield a new, very high accuracy recognition system for recognizing speech in all environments and in all languages.

For this document I use the following abbreviations: Conventional Acoustic Speech Recognition is **CASR**, and Non Acoustic Speech Recognition is **NASR**. For the basic description of the algorithms in this document, I will use the convention of identifying English words by the sound units that make up the words. They are called "phonemes" or "PLU"s. Other groupings of word sounds can be used and are discussed in texts on Conventional Acoustic Speech Recognition. There are 40 to 50 of these PLUs in the English language (depending on whose definition you use). Appendix A of this document shows a list of English PLUs, which is taken from Rabiner and Juang "Fundamentals of Speech Recognition" Prentice-Hall, 1993. Later in this document, I describe other algorithms that use combinations of PLUs (e.g. diphones, triphones, etc.) and those that use whole words, especially for use in specialized vocabularies. The algorithms in this document are not restricted to the English language, however for convenience, the examples chosen and the PLUs used are for English.

Once sensor signatures are identified (e.g. acoustic microphone outputs or radar outputs) that correspond to a given PLU with high accuracy, there are several procedures discussed in the literature for finding the words and phrases made from sequences of identified PLUs. Two important procedures that are well known to experts in the speech recognition community are the "Statistical Pattern Recognition Model" and the "Hidden Markov Model". See reference 1, and references therein for further descriptions of these techniques. The algorithms in this paper show how to prepare EM sensor data in such a fashion (i.e. in vector form) that one or the other of the above mentioned commonly used procedures (i.e. models) can be used to identify the words and sentences to complete the recognition process.

## I A: NON ACOUSTIC SPEECH RECOGNITION ALGORITHMS:

The Non Acoustic Speech Recognition (NASR) algorithms described in this paper are classified in generic groups from the simplest to the most complex. The simplest generic algorithm described in this document is used to associate electromagnetically detected single organ motions (not positions) with simultaneously detected acoustic signals. I then show how this algorithm provides background noise suppression, how it determines whether a sound is voiced or non-voiced, how it determines the rate of speech-unit delivery, the pitch of speech sound, and how it provides "on-set of speech" timing. In contrast, the most complex generic category of the algorithms described in this document involve many different EM sensors and the sensors are positioned in many locations around the head and neck (multi-position including side of face, etc.) operating on several wavelengths and phase relationships. The example I use of the "complex" algorithm describes a procedure for using three simultaneously working electromagnetic speech organ detectors, two of them obtaining many ( e.g. 30) positions of each individual organ during each 15 milli-second speech sound-formation period. As a result, in conjunction with simultaneously detected acoustic speech, this "complex" algorithm provides detailed position and motion information on all of the major speech organs as each speech sound is spoken. As a consequence, very high accuracy word-unit identifications can be generated. We believe that eventually such "complex" algorithms can "speech recognize" without the need for simultaneous acoustic speech information, although their use in conjunction with acoustic speech information will certainly be their first application. In addition, it is clear that the detailed physiological motions associated with common words provides a text independent means of unique speaker identification.

Two other generic types of non-acoustic speech recognition algorithms are described in this document. They fall in categories of intermediate complexity. They are "vocal tract model" algorithms and "word signature" algorithms. These algorithms can be used separately or can be used in conjunction with elements of the other algorithms (or of each other) to optimize the recognition of spoken speech for the application and the equipment at hand. Several other algorithms appear possible to develop and will be described at a later date. Two examples are PLU-pair algorithms that use both relative positions and relative velocities to provide unique PLU-pair identification (of which 800 are commonly used in English) and another makes use of variable wavelength and/or variable phase algorithms to establish the presence or absence of tissue "contacts" or cavity dimensional changes which characterized certain sounds quite exactly. The important elements of all algorithms mentioned in this document, including the two just mentioned, have been demonstrated. Specific examples are shown for the four algorithms described in this document.

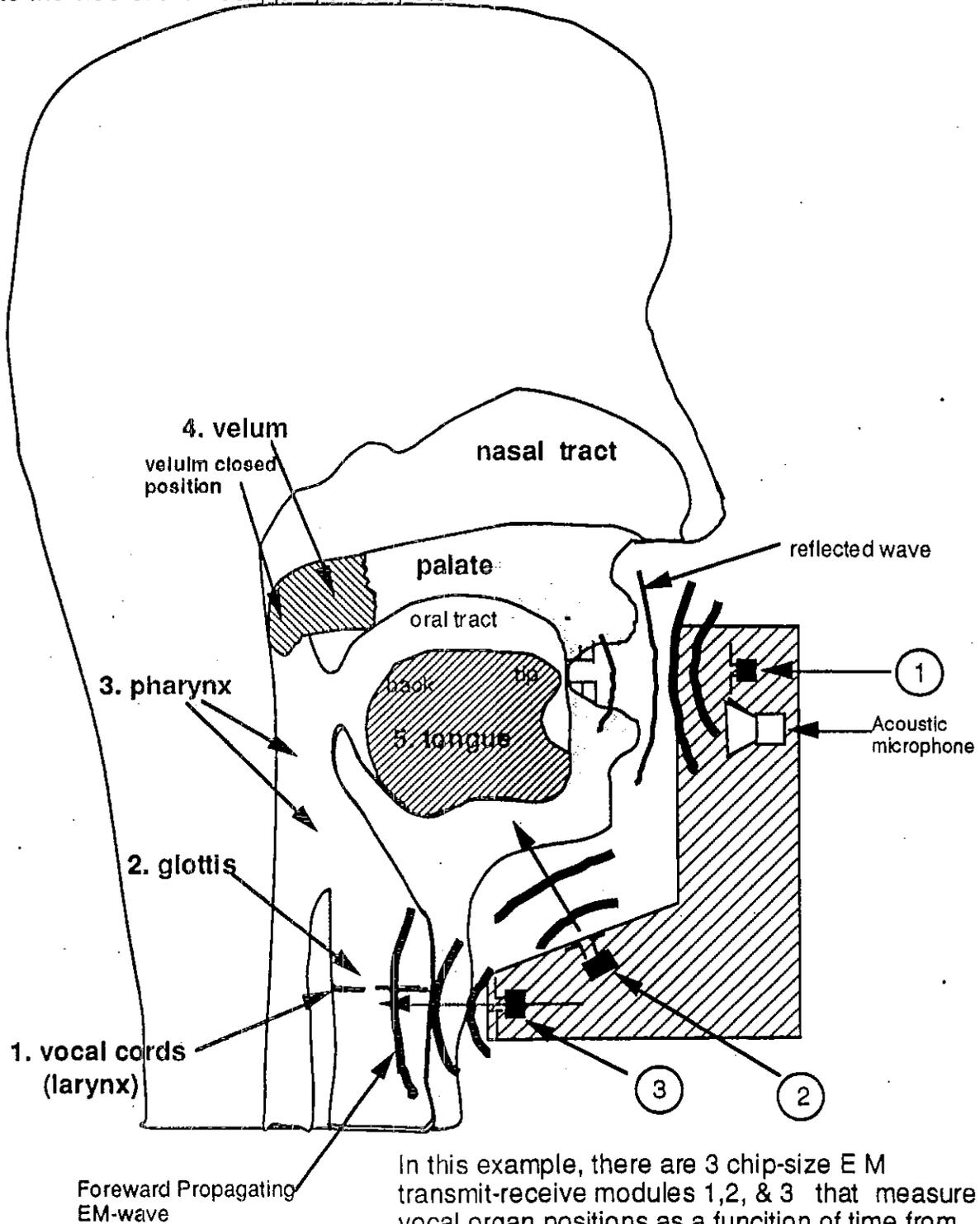
The above described generic families of real time "speech recognition" algorithms were invented for the purpose of processing information outputs from non-acoustic electromagnetic transmit-receive units (e.g. radars) which report on the positions and/or motions of human speech organs as acoustic speech is spoken. The association of speech organ locations with language sounds is not new, and is described in detail in the book "Acoustics of American Speech" by Olive et al. (Springer 1993) which is reference 2 of this document. There is a long history of using mechanical, optical (e.g. TV imaging), electrical, magnetic, and x-ray devices for observing the motions of the human articulator system while speech is occurring. An early reference to optical measurements of the vocal folds opening and closing and to x-ray imaging is given in reference 4, by J.L. Flanagan "Speech Analysis, Synthesis, and Perception" 1965. More recently, using moving x-ray images, many authors, ( see for example Pacun in reference 5) have shown that tongue and jaw positions correlate well with spoken speech and have used the knowledge to generate improved acoustic speech recognition algorithms. While these techniques have been of enormous importance in understanding the human vocal system, they are intrusive, non-portable, and dangerous unless used by experts; therefore none are very useful for real time speech recognition.

## I B ELEMENTS OF EM WAVE TRANSMISSION

Electromagnetic transmit and receive modules (radar sensor units) that are optimized for speech recognition are low power for safety (<microwatts of average power), provide centimeter to millimeter resolution, propagate through human body tissue ( e.g. at 2 GHz frequency), can be optimized by varying wavelengths for resonant structure sensing (e.g. 0.5 GHz to >4 GHz), and can operate in FCC approved bands and at approved power levels. A sensor module consists, usually, of a transmitter and a receiver located on a single chip or on a small circuit board. The received signals are sampled, averaged, quantized, digitized, phase compared, and low-pass filtered. This "on-chip" processing is done for phase stability or timing reasons, for economy reasons, and to reduce the information bandwidth for ease of transmission of information to a separate processing unit. In an on on-chip processor may be used with all or parts of the algorithms described in this memo to preprocess and further simplify the data before it is sent on. A separate digital processor, hardwired or software controlled, would use algorithms described in this document, for further digital processing to reach the desired speech recognized information goal.

The process begins by the EM transmitter sending one or several electromagnetic waves from an antenna into the head or neck where the waves reflect off of vocal organ interfaces (tissue-air or bone-tissue dielectric discontinuities, etc.) and which are then detected by receiver units. See reference 6 by T.E. McEwan. The EM probe signals in the experiments referenced in this paper were sent out every 0.5 microseconds (e.g. 2 MHz, but the rate can be changed by over a factor of 10 for more or less rapid sampling). The reflected signal is received every 0.5 microsecond (delayed by the round trip transit time of 1 to several nanoseconds from the transmitter to the organ and back). It is averaged, sampled, digitized, and temporarily stored in memory for subsequent numerical processing as governed by the algorithm being used in the processing unit. The format of the transmitted pulses, the sampling, the phase and wavelength variations (if used) and other controlled functions are governed by a hard-wired or software-controlled "control-unit". The example above yields 2000 samples of organ motion information each 1 millisecond. If we average the first 1000 values for good statistical reflection values, and then the next 1000 values -- we obtain 2 values per millisecond. A typical speech epic lasts 10 to 30 milliseconds (vocal folds excepted), thus we have 20 to 60 measures of vocal organ conditions per sensor (per epic) for our algorithm optimization. Vocal fold on-off epics are 4 to 8 milliseconds.

Cut-away view of speech organs showing typical locations of radar module and the directions of the transmitted and reflected EM waves. Other locations of the radar units are expected to be very useful, such as to the side of the neck, jaw, cheek, etc.



In this example, there are 3 chip-size E M transmit-receive modules 1,2, & 3 that measure vocal organ positions as a function of time from three views. This provides accurate position and velocity information from all speech organs

## I-C INFORMATION PROCESSING:

Speech organs move relatively slowly. For example, the literature shows (and we measure with radar sensors) 4-8 millisecond cycles for vocal cords and 10-30 millisecond cycles (or longer) for tongue, jaw, throat, lip, and velum muscles. For most of the examples in this document we will use a 15 millisecond time as the articulator time or the "speech-epic" being considered; however each organ sensor would in practice be optimized for the desired information acquisition, such as the 5 millisecond cycles for the vocal fold open/close times of a female speaker. In the typical speech-epic, up to 30,000 radar samples can be taken by each EM sensor unit (when sending at 2 MHz) of each organ system before the organs change shape to start forming a new sound. As the 30,000 pulses are received a wide variety of processing can take place, depending upon the application. Procedures commonly used in our demonstrations include the averaging of a 1000 received pulses for each position in a time step within the sensor itself. Experiments show that 1000 received signals provides sufficient signal to noise to obtain a given organ's condition, thus one has time to perform 30 such averages within the time that an organ moves to a new position. For example, we have moved a range gate to present different reflection positions, each a few millimeters apart, from one organ (jaw and tongue) during one period of observation. In Section V of this document where a multi-organ algorithm is discussed, I show an example of how the processing is accomplished. During the available time between organ motions other measurements can be made in addition to moving a range gate; two examples are: changing transmit wavelength to detect resonance effects (e.g. tongue touching palate experiment produces giant resonance effect) and changing the phase of a multiple wave transmission in a homodyne detector to detect movements of nulls and resonance's.

After the data is taken for a given period (e.g. 15 msec) further averaging and normalization of the data from each time step can be done. Time adjustment of distance positions (i.e. round trip reflection times) to a known fiducial such as the front of the face can be done so that all data sets, as time progresses and the sensors or the head moves, are commonly referenced. The fitting of the pattern over several 15 millisecond speech epics to a model of vocal organ time motions, such as LPC models, can be done. Fourier transforming of the data into frequency or position space for ease of identification of characteristic periods such as vocal fold on-off frequencies for pitch information can be done. The subtraction of patterns gathered from a previous time epic (i.e. subtracted from the present step) can be used to generate motion information of the organ interfaces. Finally, we form the information from a specific speech unit into a pattern that can be matched against the PLU-EM sensor patterns which are stored in a known digital library (also called code book) of speech element patterns. When a match of the radar pattern to a

known pattern leads to a sound unit or PLU recognition, I define the algorithm as having rendered the spoken speech accurately (in real time) into a computerized or printed symbol. Further post processing is required to correct word spellings, sentence structure, identically spoken words (e.g. to, too, or two) and other forms of information associated with the concept of automatic speech recognition, but these are not described in this document.

#### I-D RECOGNITION:

In the simplest algorithm described in this document, a comparison is made between the output from the radar speech recognition unit and a more conventional acoustic speech recognition unit. If both systems have high probabilities of their respective identification and they agree on the identification, then the identified word will be more accurately rendered than if just one system were working alone because the joint probability of identification is higher. For example, present acoustic recognizers have at least 5% error rates. Presently used EM sensors have estimated error rates for detection of speech organ motions that are 5% to less than 1%, depending on the sensor and application conditions. In addition, we have found that certain PLUs or words are known to be more easily recognized by EM sensors than by acoustic speech (and vice versa). This means that the probability of identification of each sound can be associated (in the algorithm library) with a known accuracy weight depending on the sensor used. This probability can be used in the pattern matching algorithm for single or joint identification described later in this document. In summary, word recognition can be improved by appropriately weighting the output of the recognizer (EM or acoustic) that is known to be accurate for the word sound being analyzed.

#### I-E FOREIGN LANGUAGES:

The speech recognition algorithms described in this document are language independent in that the radar sensors can obtain vocal organ position and motion information from speakers using any language, (i.e. they work on English speakers as well as Chinese and other language speakers). However for each language, a non-acoustic speech recognition system's radar-sensor-suite and associated algorithms will need to be optimized for each language's specific speech organ motions. In addition, the required speech recognition post processing systems will be chosen for correct spelling, proper grammar, proper sentence structure, etc. in the relevant language. Another important application of this technology will be the teaching and learning of foreign languages, or the correction of speech problems by native speakers. The information described in this document allows one to detect speech organ misplacement, a major problem in language pronunciation.

## II. EM-SENSOR SIGNAL DESCRIPTIONS FOR ALGORITHMIC PROCESSING:

### II A. Principles of Radar Detection of Biological Tissue Positions and Motions

Electromagnetic transmit-receive modules (radars) designed for body organ motion and position detection are optimized to meet the needs of the application. For medical purposes, McEwan (6) has invented and optimized specialized micro power radar sensors for the purposes of heart and lung motion diagnostics. For our speech recognition experiments, similar radar sensors were modified. For commercial speech recognition purposes, it is expected that LLNL's present micro power radars will be further optimized. It is also expected that others will design new systems to optimize the acquisition of the information needed from the speech organ being sensed and for the algorithm used to convert organ information into recognized speech information. Radar system parameters that are optimized include the EM wave transmission properties from the antenna, through the air into the head or neck tissue, reflection from the tissue interfaces (e.g. tissue-air, tissue-bone, tissue-teeth, etc.), transmission through the tissue, and back propagation to the receiver antenna. Additionally, resolution requirements on interface positions lead to transmit and received EM-wave time-resolution and wave-phase requirements. Also "speckle" issues pertaining to coherent interferences between scattered signals from multiple locations and from multiple parts of the transmitted wave need to be resolved, as well as receiver signal to noise requirements and repetitiveness of pulse transmissions need to be defined. Additional considerations such as relative interface locations and vocal cavity dimensions, and locations in the surrounding tissues influence the selection of the radar's wavelength(s). Finally, safety requirements and user sensitivities, such as FCC regulations limiting the power and the frequencies available for use, define the design of EM sensors for speech information gathering.

Most experiments conducted by the author and T. McEwan used 2.0 GHz pulsed radars in a variety of transmit and receive modes. They transmit less than 1 microwatt of average power which is 10 to 1000 fold lower than all accepted (US and international) safe continuous rf and microwave power exposure levels. While this frequency is convenient because of the availability of experimental equipment, and it has been used to demonstrate the elements of the algorithms described in this document, the 2.0 GHz frequency is not optimal for all speech organ information gathering requirements. I provide a short description of radars optimized for speech recognition applications.

## II B. Some Radar Electromagnetic Wave Properties:

The wavelength  $\lambda = c / \nu$  is assumed (for this document) to be the transmitted wavelength from the radar transmitter. It is defined by the time period,  $\tau$ , of a single wavelength or by the period of a half wavelength,  $\tau / 2$ , which define the frequency  $\nu = 1 / \tau$ . The wavelength is  $\lambda = c / \nu = c\tau$ . Issues associated with definitions of the wavelengths, frequencies, and bandwidths of single (i.e. impulse) 1/2-wave-pulse transmissions, of gated wave packages, of multiple waves, of chirped waves, of phase varied waves, etc. are described in texts on radar. For example, see reference 7 by M. Skolnik "Radar Handbook". I will not discuss these definitions in any detail because they do not strongly influence the algorithm descriptions in this document. For most of the examples in this document, I use  $\nu = 2.0$  GHz which leads to:

$$\lambda_{\text{air}} = (3 \times 10^{10} \text{ cm/sec}) / (2.0 \times 10^9 \text{ Hertz}) = 15 \text{ centimeters in air}$$

$$\lambda_{\text{water}} = \lambda_{\text{air}} / \sqrt{\epsilon_w} = 15 \text{ centimeters} / \sqrt{64} \approx 2 \text{ centimeters in water}$$

where I approximate the dielectric constant of water to be 64

$$\lambda_{\text{tissue}} = \lambda_{\text{air}} / \sqrt{\epsilon_t} = 15 \text{ centimeters} / \sqrt{25} \approx 3 \text{ centimeters in tissue}$$

where I approximate the dielectric constant of muscle tissue to be 25

Another important property of EM wave propagation is the diverging of a transmitted wave as it leaves an antenna and as it propagates from one dielectric medium into another. These issues are discussed in detail in the above mentioned book on Radar and in any text on EM wave theory. In particular, a transmitted wave from an antenna diverges from an antenna with a full angle  $\theta$ , where  $\theta \approx \lambda / d$ , and  $d$  is the dimension of the antenna. Thus a wave leaving a dipole antenna that is 1/2 wave in total dimension diverges very rapidly and, if positioned far from a head, the wave would be much larger than the organ or organ interface targeted for measurement. As a result, little energy will reflect from the organ. In addition, at the interfaces between air and a high dielectric material such as water filled tissue, strong reflections occur with typically only 10% of the wave being transmitted into the high dielectric constant (i.e.  $\epsilon_t$ ) medium. However once inside the tissue, the EM wave divergence rate is reduced by the  $\sqrt{\epsilon_t}$  (i.e. the index of refraction) and the wave diverges much more slowly. Its propagation is governed by the principles of physical optics which include diffraction, scattering, constructive and destructive interferences, and related phenomena which take place in the complex structures inside the head. Since many of the structures of the speech organs and cavities are of the dimensions of the EM waves used (e.g. 3 cm in tissue), very strong resonant propagation effects can occur and have been measured by us. By changing  $\lambda$  one can optimize the detection of desired resonance responses. These form the basis for a class of detection algorithms based upon resonances or "standing wave" properties to

enhance or simplify the information gathering properties of these systems. These effects are also strongly dependent upon the polarization of the EM wave being transmitted. Several techniques are used to optimize the wave path for the measurements described in this document. They include the use of shorter wave transmitters (consistent with propagation through tissue), placement of the antenna close to the head or jaw before divergence has developed, the use of dielectric focusing or ducting materials (e.g. high  $\epsilon$  foams), the use of multiple-wave length antennas for focusing and narrower beam formation, the use of quadrature techniques (6), and the selection of the EM polarization and the angles of incidence of transmission into the throat and head.

### II C. Transmission and Reflection of EM wave from Vocal Box in Neck

The sketch below in Fig. II-1 illustrates the transmission and reflection of a single linearly polarized EM wave pulse in a number of locations, each one wavelength apart. An important observation is that the wavelength of the wave as it enters the neck shortens and the propagation speed "c" slows down by  $1/\sqrt{\epsilon}$ . This occurs because the tissue is a material with dielectric constant  $\epsilon$  greater than  $\epsilon_0=1$  for air. In addition, the amplitude of the electric field drops for two reasons. The first is that a significant fraction of the forward propagating EM wave reflects at the first surface of the air skin interface, and the second reason is that in a dielectric medium, the E field drops because of the high dielectric constant. The shortening and slowing of the wave makes it possible to measure the size and location of structures internal to the head that are a small fraction of each radar pulse length in air. Since it is common practice to measure distances to less than 1/10th of the EM pulse length dimension, we can see structures that are 1/10 of a half wave pulse of 1.5 cm of the wavelength in the tissue, or 1.5 mm. By seeing we mean that it is easy to detect changes in the EM wave reflectivity associated with the absence or presence of 1.5 mm structures, or more importantly, we can easily detect changes in the positions (interfaces) of vocal organ structures down a 1.5 mm scale as speech organ motion occurs between one set of pulses to another set of pulses after the motion has occurred. These position changes are associated with motions of the vocal organs as they form themselves to prepare for the next speech sound.

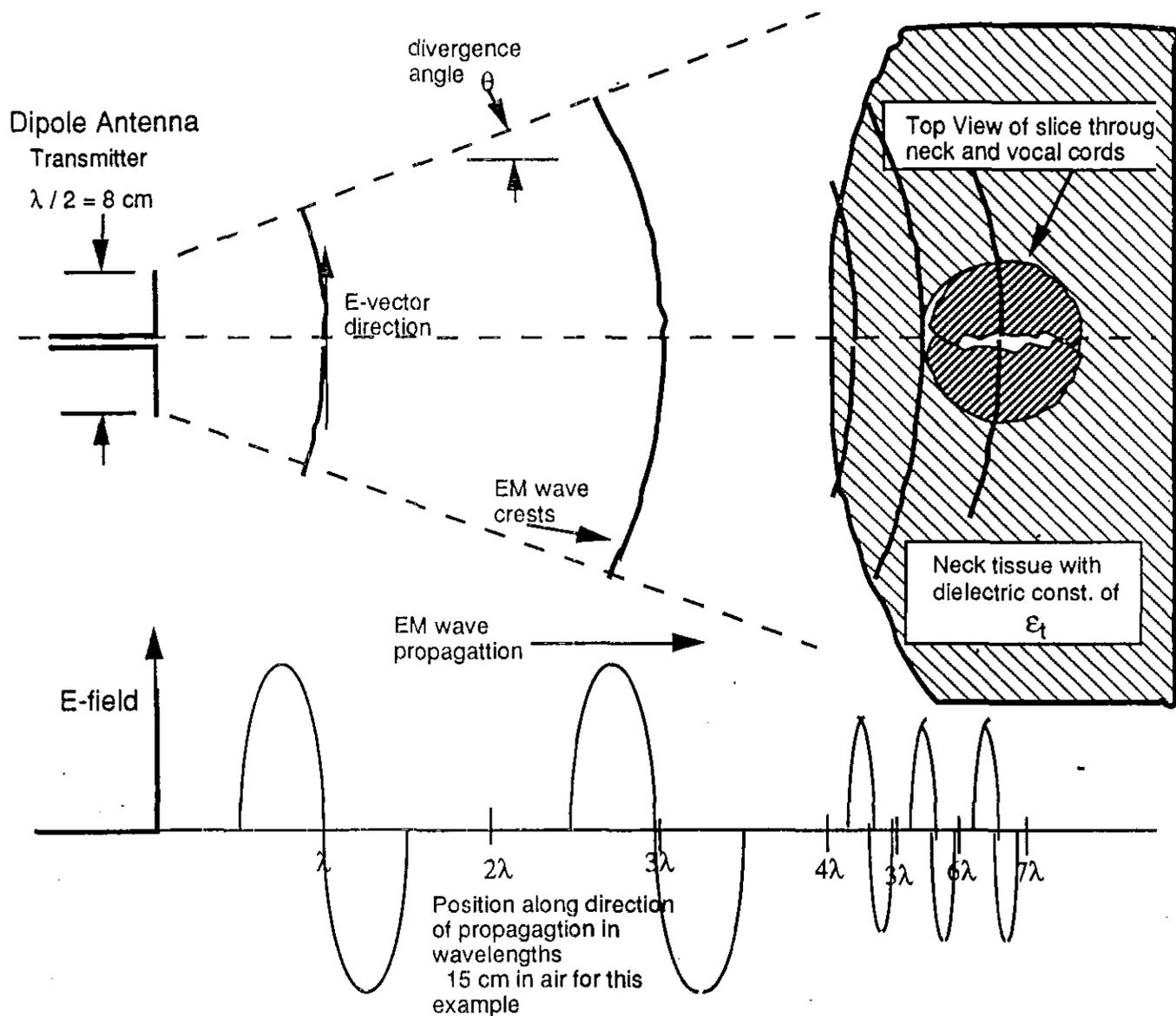


Fig. II-1 Schematic of a single EM  $1/2$ -sine pulse being transmitted in the direction of the vocal cords in the neck as time increases. The drawing only shows forward propagating pulses, reflected pulses are discussed later. The EM wave pulses slow in velocity as they enter the dielectric medium of the neck, and begin to encounter the air tract (Larynx) and the vocal folds (cords). Notice that the wave diverges rapidly from the simple dipole antenna and most of the power misses the vocal cords. In actual practice the antenna is placed closer to the skin, and dielectric "matching" or focusing materials can be used to improve the direction and efficiency of coupling into the organ being measured. Also note that the EM field is polarized in this example; the scattering is sensitive to polarization angle and the reflection from interface is a strong function of polarization angle.

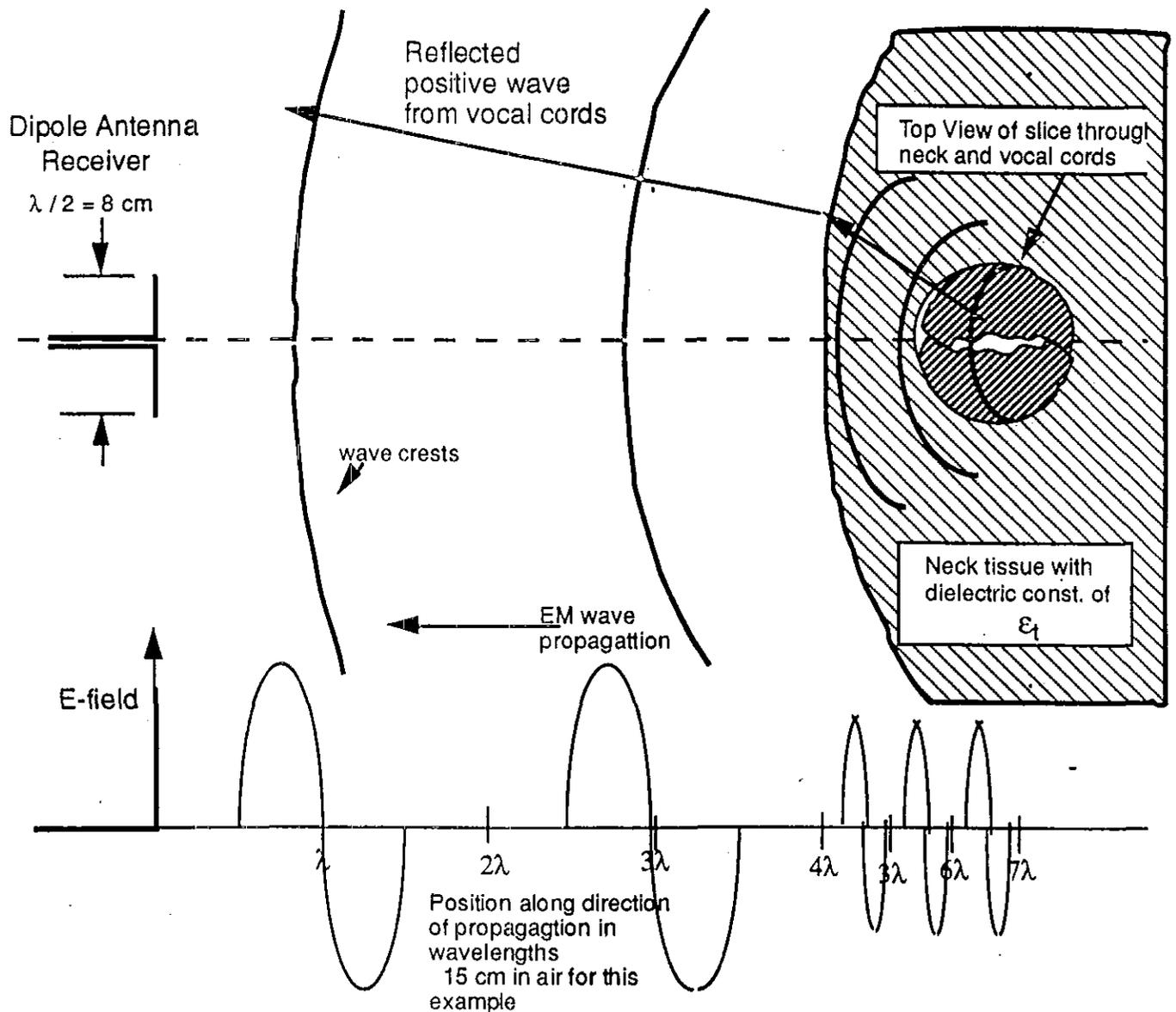
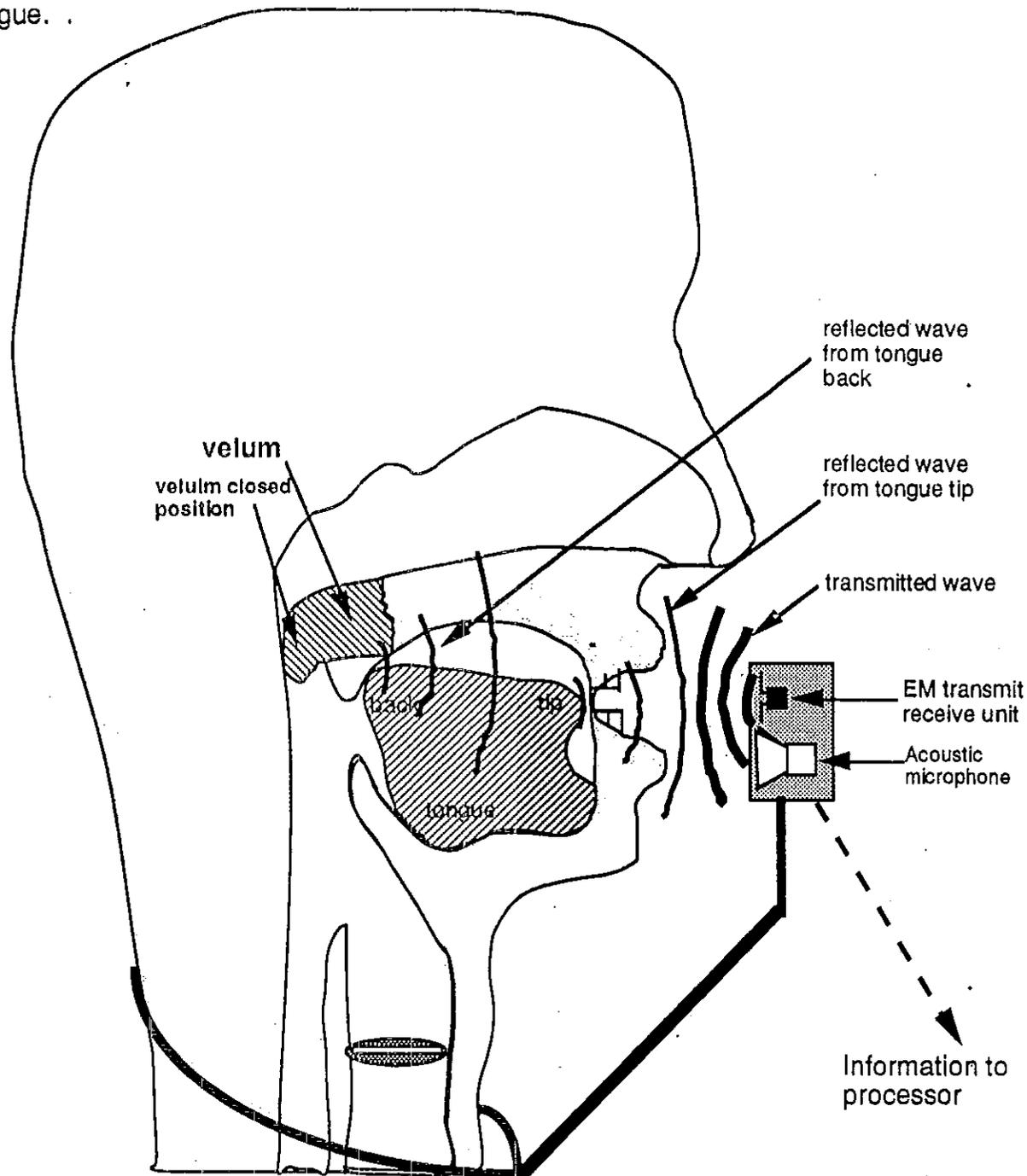


Fig. II-2 Single EM wave being reflected from vocal folds in neck in many locations as time increases and propagation toward the receive antenna. The sketch does not show the scattered EM waves from the air-skin interfaces which are quite strong. The air-skin interface reflection is easily filtered because it doesn't move during a word sound, and the vocal folds do. Designers use filters to detect faster intensity changes that are associated with position changes in the vocal folds. For example if the vocal folds move nearer to the antenna, a more intense signal is received or if they move further away, a less intense signal is received. Unwanted reflected signals can also be filtered out by noting that the skin reflection returns earlier to the antenna than the reflection from the vocal folds which are three wavelengths into the neck. A range gate can reject pulses which return too early, and accept those parts of the wave (e.g. positive part) that return on time.

#### II D. EM wave Reception and Processing:

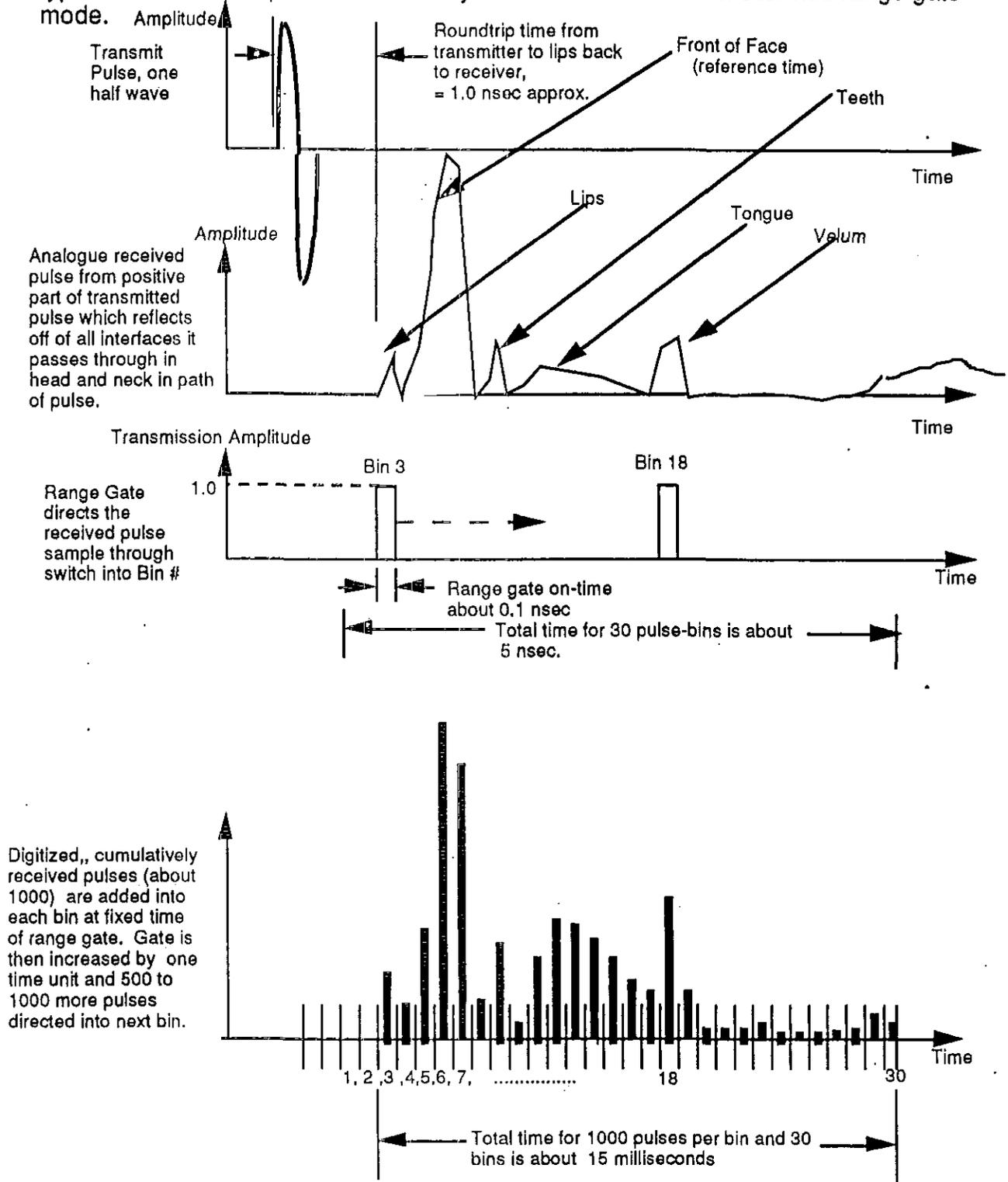
After the radar pulse is received it must be processed, correlated with other pulses from other organs and correlated with acoustic data and fed to an algorithm which automatically selects a word-unit (i.e. PLU), and displays it to the user or customer. Because these radar sensors easily generate approximately 2000 pulses each millisecond (our experiments have been done at 2 MHz transmit rates), one can average 1000 pulses for each reflected range position or for each wavelength, etc., then one can change to a new range, wavelength, etc. and average another 1000 pulses, and thereby measure up to 30 parameters during each sound formation epic of 15 milliseconds. The Figures II-3 and II-4 (see below) show conceptually simple ways of measuring the locations of all of the organ interfaces from the front of a face through the back of the throat. The locations at a given time can be associated with a given speech PLU and by knowing all of the organ locations for each sound epic ( e.g. about each 15 millisecond window except for vocal folds which open and close each 3 milliseconds) one can identify the sound being spoken. By associating the PLUs identified using the EM sensors (NASR) and comparing with the PLUs identified by conventional acoustic techniques (CASR), one has a very high probability of identifying the correct PLU. Our data and calculations (so far) indicate greater than 95% accuracy for all NASRs organ measurements (and > 99% for certain on-off measurements). Experiments at LLNL and elsewhere indicate 95% or greater for CASRS. We estimate that the joint recognition statistics of both systems together will lead to speech recognition accuracy-error rates of less than 0.1%. This is highly accurate word recognition, approaching human hearing standards. However, as discussed below and later in this document there are many speech recognition situations where complete organ location and motion information is not necessary. Their conditions are described in several of the algorithmic descriptions later in this document. In addition, there are several other algorithms which use the special information available from the EM sensors to provide new ways of recognizing speech in both specialized and generalized situations. They are described later in this document as word signature algorithms and as motion pattern algorithms.

FIG. II-3: Cut-away view of speech organs showing location of a range gated radar module directing its transmitted beam into the mouth horizontally and a reflected EM wave from the back of the tongue.



In this example, there is one chip-size E M transmit-receive module that measures face front, lips, teeth, tongue parts, velum, and pharynx conditions as a function of time from a horizontal view. This provides up to 30 or more accurate position locations during each speech epic. By subtracting the locations obtained during one epic from the next epic, and dividing by the time between epics, one obtains horizontal velocity component information from all speech organs.

FIG. II-4 Short transmit pulse and scanned range gate configuration, showing typical received and processed data by horizontal EM sensor in scanned range gate mode.



We have obtained very useful information on speech organ motions by measuring whole organ motions. This type of very simple, yet informative, information on whether an organ has moved from one speech epic to another can be understood by considering the range gate in the Fig. II-4 above to be about 5 nanoseconds in duration so all of the reflected signal signals from all of the horizontal interfaces are averaged over 30,000 times, digitized and stored into one bin. This processing algorithm causes all interface data from one horizontal mouth configuration to be averaged over one speech epic and stored in the first bin (#1) location. Then the process is repeated and another speech epic is averaged and stored in the next bin 2. By comparing one bin number to the next (e.g. by subtracting or by ratioing one to the next) one finds a correlation that is a consequence of the different total path reflection of the single transmitted wave with all of the speech organ interfaces associated with the formation of a given sound in the 15 millisecond epic being measured. The next mouth configuration, measured one or more speech epics later, gives a different average-reflected EM signal value because there is a different cumulative reflection of EM waves from the new set of speech organ interface locations. For example, this technique was first used by Holzrichter and McEwan to measure the motions of vocal folds (i.e. vocal cords) as they "burst" open and then close thereby exciting the vocal resonator system. This technique provides very useful information on the presence of "voiced" (i.e. vocal fold active) sounds, which with the proper algorithms, enables the simple discrimination of similar sounding acoustic sounds units such as "b" and "p", the first voiced (i.e. vocal folds moving) and the second not voiced (i.e. vocal folds still).

There are additional combinations of transmitting EM pulses for the speech recognition application. They include transmitting multiple pulses (i.e. several waves of a pulse train over a given duration such as 3 nsec.) and using a short duration, medium duration, or a completely open range gate to collect the reflection EM waves. Another important configuration uses the above configuration in Fig. II-3, but compares the reflected EM waves to a part of the transmitted wave (suitably delayed) in such a configuration (homodyne) that both the phase and amplitude change between the reflections and the fixed transmitter phase (from speech-epic to speech-epic) is detected and used for signal processing. Other techniques include changing the transmitted wavelength from sample to sample during a given speech epic to detect resonance effects in the reflections from the vocal cavities and organ configurations. Another technique is to move the phase of transmitted single or multiple pulses relative to a fixed transmit gate and to a fixed receive range gate and by using a homodyne mode one can interferometrically measure distances from the antennas to the speech organ interfaces and back. Many conventional radar techniques and several new techniques can be used to obtain speech organ information, which when processed by appropriate algorithms, can provide very valuable information for accurate, economical, and rapid speech recognition.

All of the algorithms described later in this document use the property of obtaining speech organ position or motion information through transmitting and reflecting EM waves from the speech organs. The basic ideas of these new ways of processing the EM information, often in conjunction with simultaneous acoustic information, are described in this document as algorithmic procedures. In addition, new algorithms built from the basic building blocks (i.e. procedures) described in this document can be devised and applied to specific applications.

### III. SINGLE ORGAN NON-ACOUSTIC SPEECH RECOGNITION ALGORITHMS

#### III Introduction:

The actions of single speech organs can guide important decisions made by traditional acoustic speech recognition systems, however it is not possible to use the non-acoustic (radar) signature from a single speech organ motion to uniquely identify a word-sound. Thus these algorithms are used primarily in the joint speech recognition mode ( where EM plus Acoustic recognizer sensor & algorithms are used together). However, as discussed in Section V., several such single organ sensors and algorithms can be combined into a multi-organ, multi-location system for either complete non acoustic recognition or for very accurate joint non acoustic and acoustic recognition systems. Single organ motions provide important information that aid conventional acoustic speech recognition algorithms to provide more accurate, faster, and more economical overall speech recognition. In addition, single organ information is the key to speech recognition (and synthesis) based upon vocal tract recognizer models, which are described in Section IV. By single organ motion we mean signals associated with the organ moving as a whole (e.g. vocal folds opening and closing or resting); we also include methods using time differentiation of organ motion (e.g. the tongue tip moving at rates faster than the tongue body), and we include methods where during a given sound epic (e.g 15 milliseconds) several measurements are made of specific organ-part locations (e.g. tongue tip vs tongue back) or several resonance effects are measured by using one or more wavelengths of the EM sensor to measure the tongue-palate dimension, or the tongue-tip to the alveolar-ridge palate contact, etc. Conventional acoustic speech recognition systems (CASRs) have problems other than speech recognition, which lead to their nominal 5% error rates in quiet office environments, but whose error rates exceed 10% when used in noisy environments, when used by stressed speakers, or when used by dialectual speakers. Several of the algorithms described here are used to cancel ambient noise effects, and to aid in determining the start and stop of speech.

Various manifestations of EM sensors were described in the introduction to this document. When optimized for the single organ-information desired, and in conjunction with single organ non-acoustic algorithms, they can provide the following information:

- 1) onset of speech time
- 2) background noise rejection
- 3) presence of voiced speech
- 4) pitch of speech
- 5) rate of speech
- 6) rhyming PLU differentiation
- 7) excitation dynamics and/or vocal tract shape changes vs time for model based recognition systems (see section IV)

Figure III-1 Shows a simple rendition of a speech onset detector, background noise suppresser, and voiced-unvoiced PLU discriminator, pitch of speech, rate of speech, and rhyming discriminator in a simple version of a non-acoustic acoustic speech recognition the system ( two example-sensors are shown, one for vocal folds, the other for tongue positions).

Inside-Head view of non Acoustic onset of speech detector, tongue motion, and vocal fold location

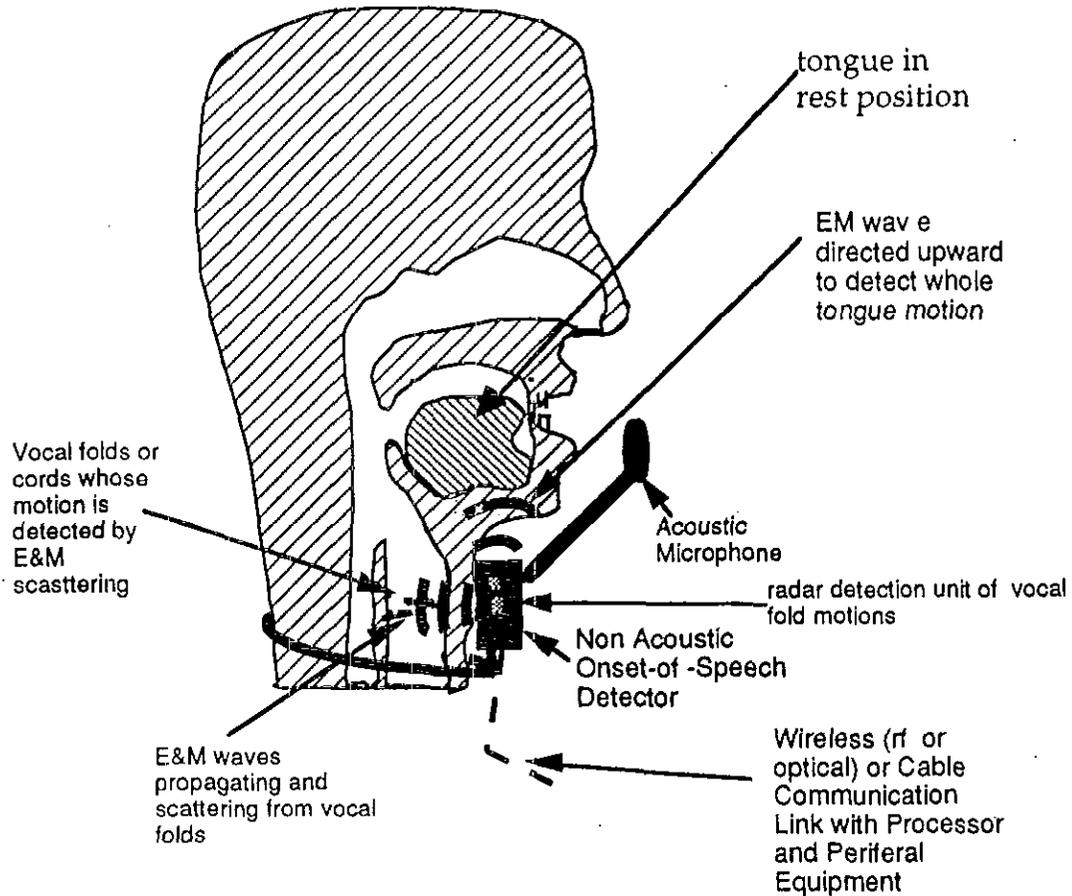
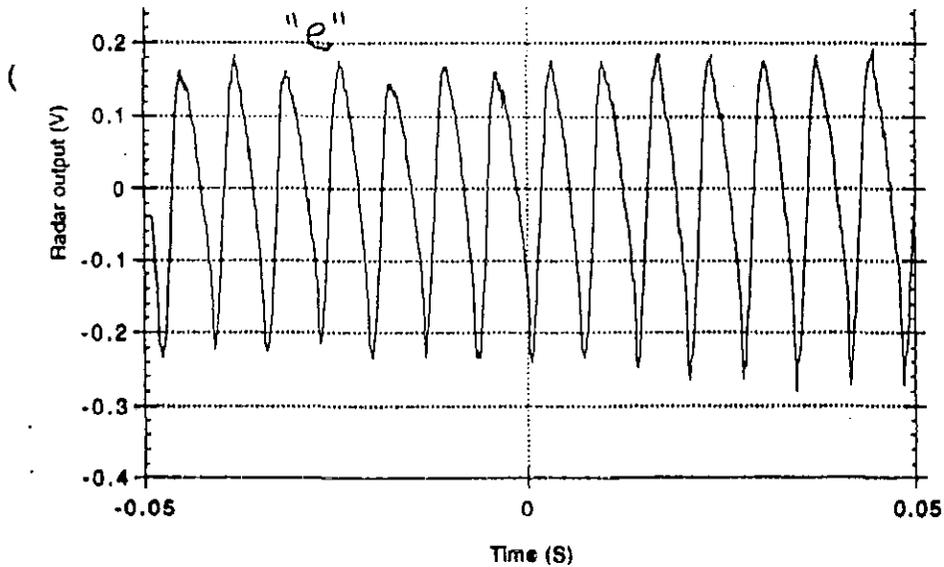


Figure III-2 (see below) shows two examples of vocal-fold radar- and acoustic speech microphone-outputs. The top figure shows processed (averaged and low pass filtered) EM reflection data from the vocal folds opening and closing while they say the sound "e". The second shows a simultaneous vocal fold and acoustic signal as the male subject says the two words with all voiced PLUs "one" and "two". These examples show the fidelity, ease of detection, and the information that is available with appropriate algorithms.

III-2a



III-2b

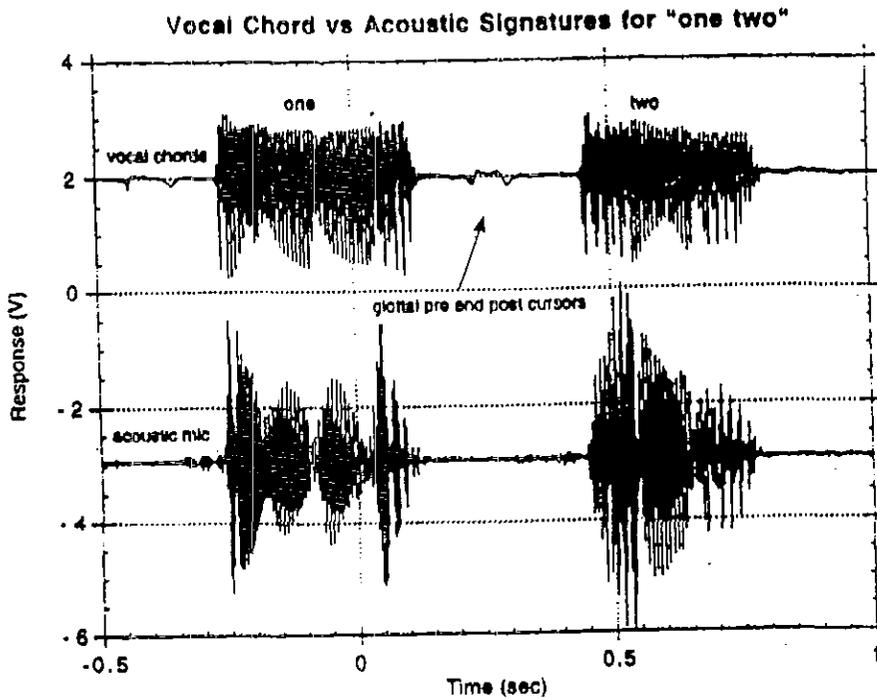
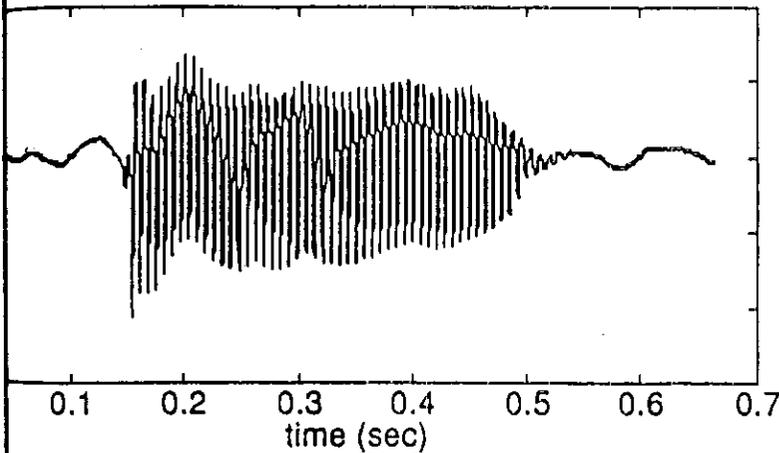
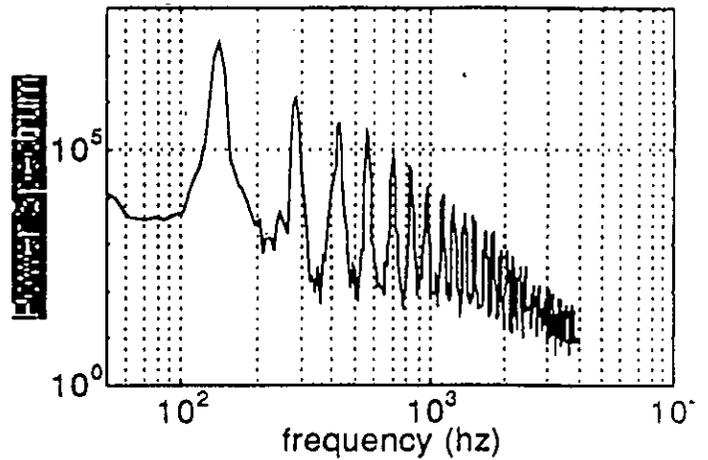


Figure III-3 ( see below) shows another example of the word "two" as a female subject says the word. In addition, it is Fourier transformed to show the characteristic frequencies associated with vocal cord motions, with acoustic sounds, and with the effect of acoustic sounds back on the vocal cord motions. While the "bursting mode" motion of the vocal folds are not affected by the influence of the acoustic resonator being excited, the small motion, higher frequency vibrations of the vocal fold membranes and surrounding membranes are "vibrated" by the presence of the acoustic waves in the vocal tract. The radar sensors are so sensitive (due to the clever filtering and noise reduction techniques) that they can pick up acoustic vibrations of the membranes as well as the forced motions of the air flow. This information can be useful for vocal-tract-model based algorithms where feedback of the air column onto the glottis is important.

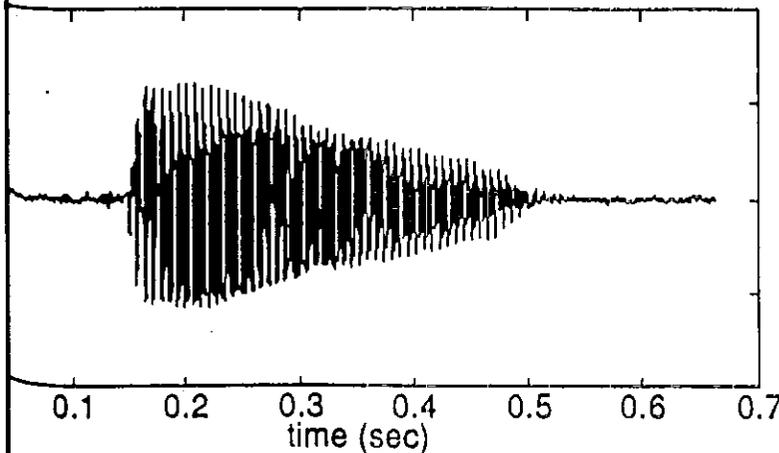
FIG III-3 :  
RF time waveform of the word (two)



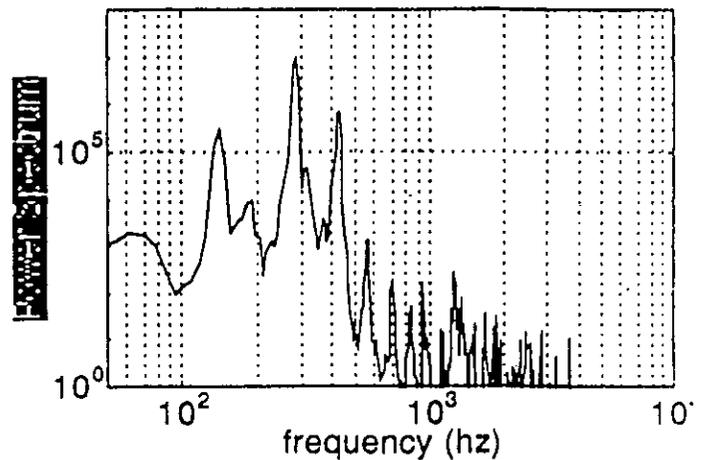
RF signal spectrum of the word (two)



Acoustic time waveform of the word (two)



Acoustics signal spectrum of the word (two)



### III. A. ONSET- AND END-OF-SPEECH ALGORITHMS

#### III A1: Vocal Fold Motion Onset Algorithm:

Vocal folds move when voiced sounds are formed. Most all English sounds are voiced and most words contain one or more voiced sounds within each second of speech. English speech usually contains about 3-7 word-sound units (PLUs) per second and a non-voiced sound is statistically always followed by a voiced sound every 1 to 2 PLU units. This property can be exploited by noticing that the radar detector Fig III-4 below shows an algorithm for detecting the onset of speech by using both an acoustic and radar unit at the same time. The issue for this algorithm is that some words begin with non-voiced sounds such as "s" in "sam", thus the algorithm must be able to back up to catch sounds missing by the onset detection by the radar. The following chart describes such an algorithm which has been testing manually.

Figure III-4a

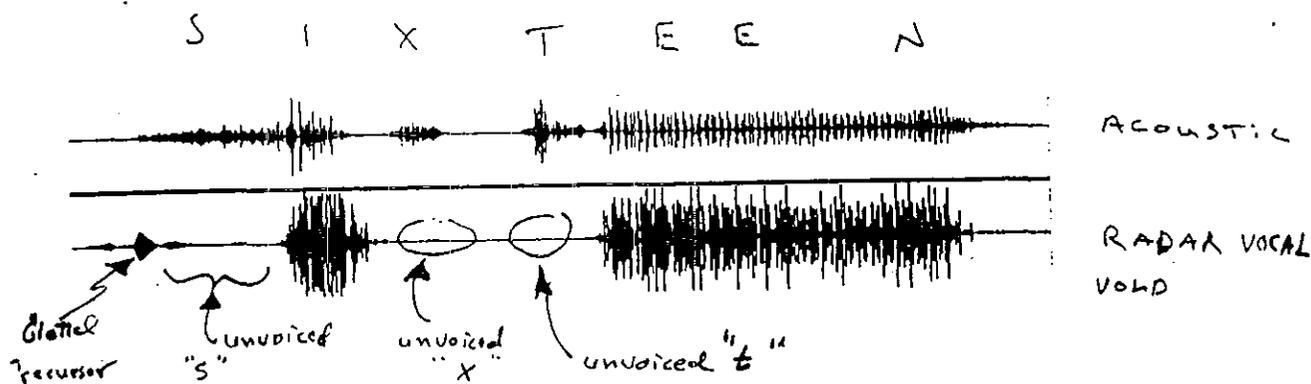


FIGURE III-4b

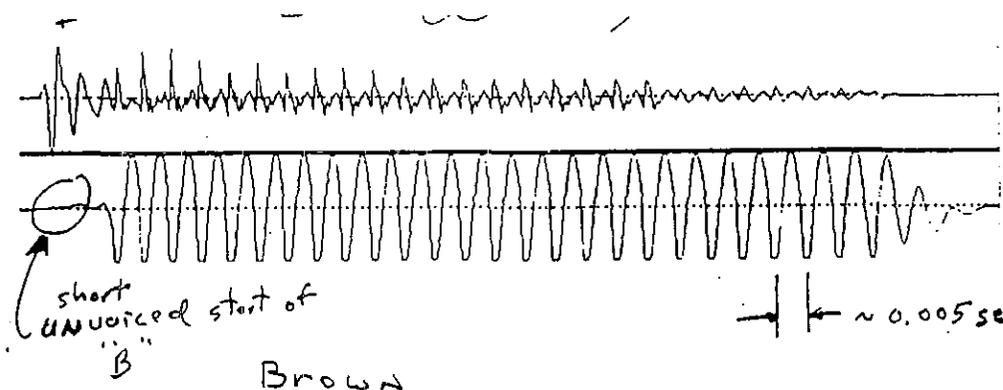
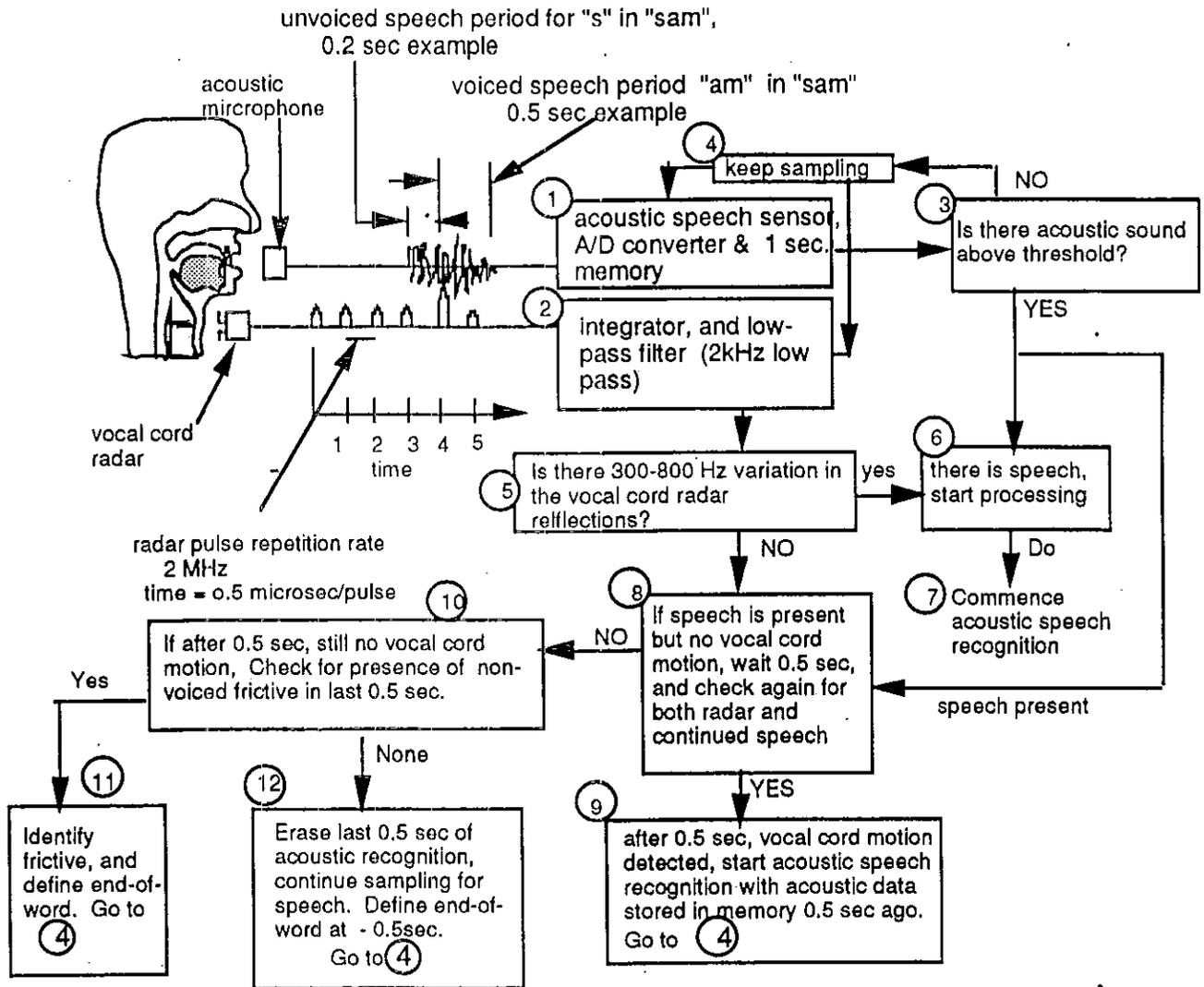
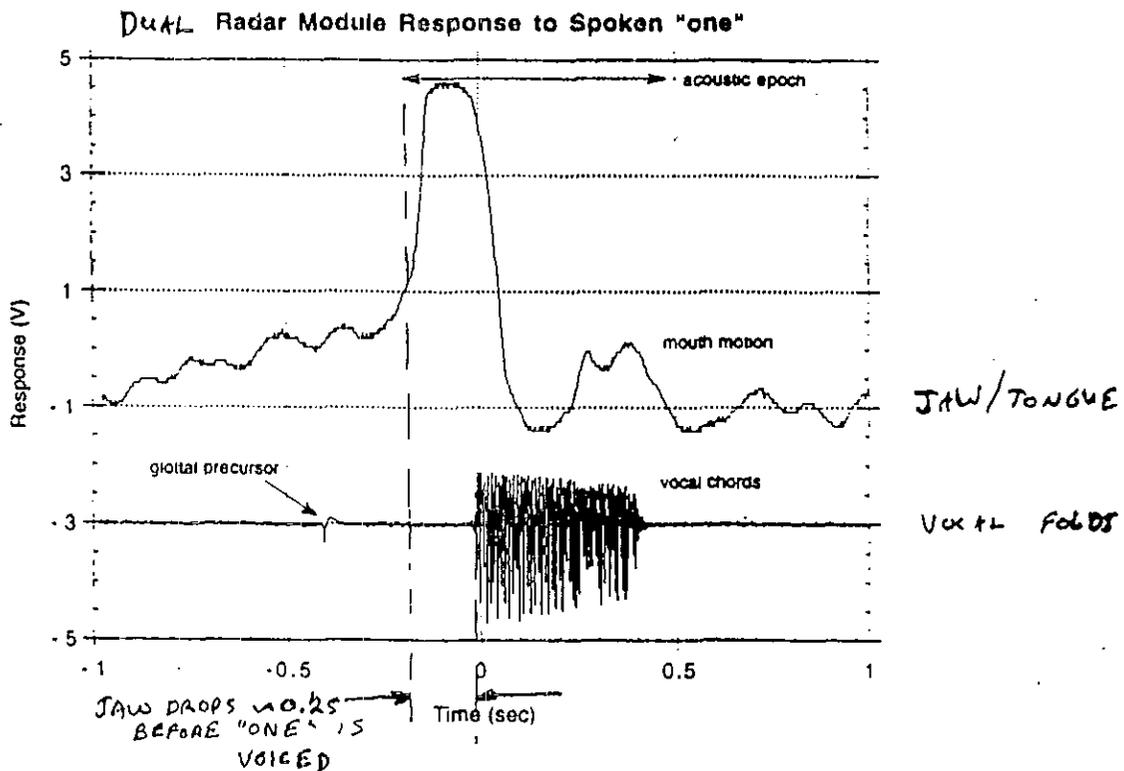


Figure III-4:  
 Flow Chart for Single Organ motion joint Acoustic non-Acoustic  
 Speech Algorithm e.g. the word "sam"



### III A2: Tongue Motion Onset/End Algorithm:

The sketch shown above in Fig. shows the radar unit pointed at the vocal cord area, and this has the same function as radar unit as number 3 unit in Fig. 1 in the introduction. However one can use micro-radar number 2 in Fig. I-1 in the introduction, and also in Fig. III-4 above, to measure tongue motion and jaw motion. Tongue and jaw motion are in some respects better indicators of the onset of speech than vocal cords because they move to start unvoiced sounds as well as voiced sounds. The same single organ arguments use to describe vocal cord motions pertain to the use of the single organ tongue (or tongue-jaw coupled motion). The only change is that the radar in Fig. III-4 is pointed upward through the underside of the jaw and the processing boxes labeled 2 and 5 have their filters and other detector constants changed to accommodate the fact that the tongue moves at a rate of less than 100Hz (much lower than the vocal folds at >250Hz). In addition, the "wait-for-speech" times of 0.5 sec in boxes 8,9, and 10 in Fig III-4 can be changed to accommodate the statistical optimization of speech start times and tongue motion for the vocabulary being used (e.g bond trading, etc.). We find experimentally, that in fact, there is often tongue and jaw motion occurring slightly before speech starts when the tongue parts move and the jaw drops from a rest position to the needed speech configuration for the speaking to begin. See Fig. III-5 below.

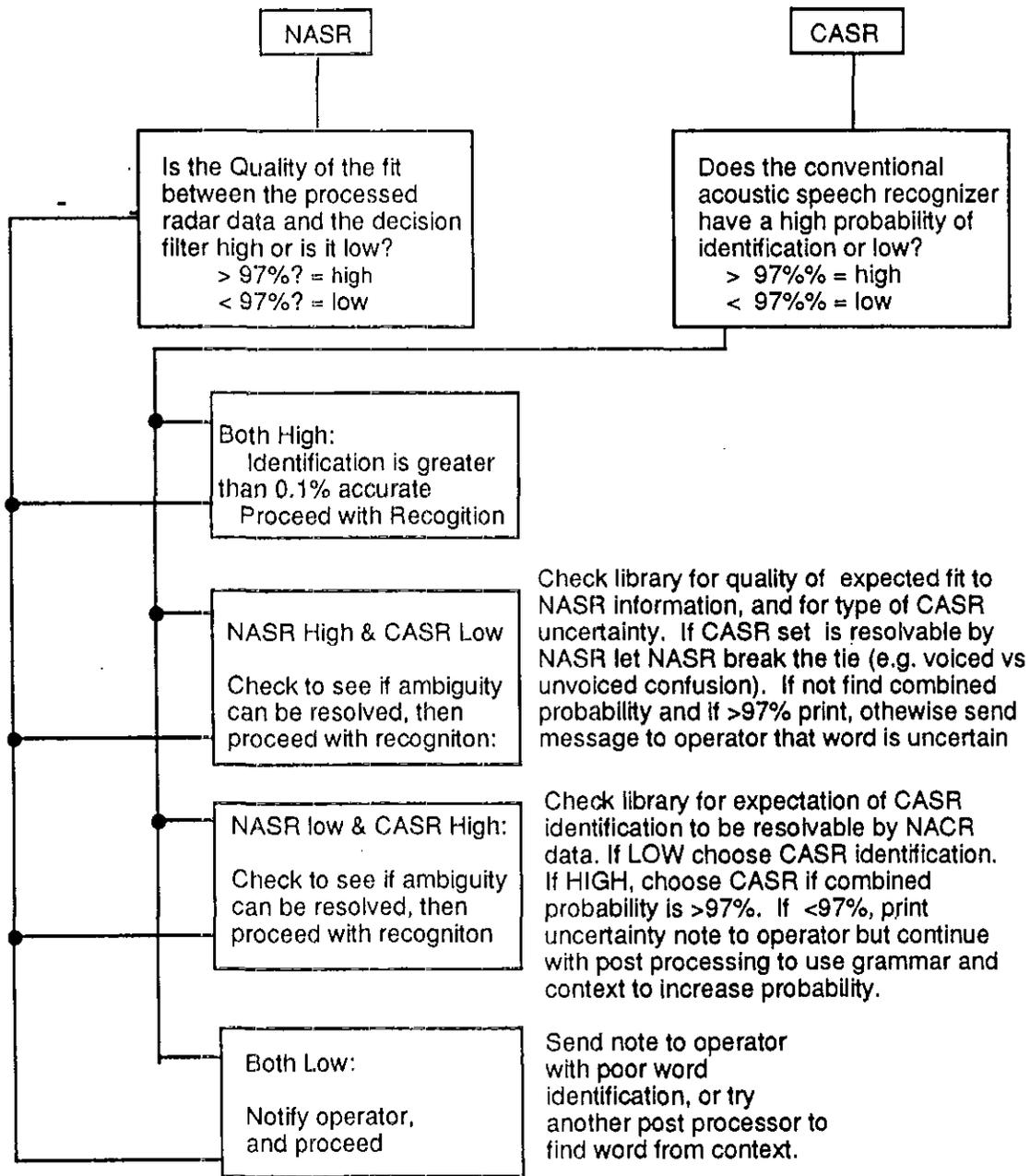


### III A3: Dual Organ Onset-of-Speech Algorithm:

If two EM transmit-receive units are used together such as the sensor units noted in Fig. III-1 above, then a vocal fold motion algorithm and a tongue motion algorithm can be used in parallel (both described directly above in IIIA1 and IIIA2). Their outputs can be joined and together with the acoustic signal, a vote can be taken to decide on the start of speech. An example that works is that if an acoustic signature plus either vocal fold or tongue/jaw is present, then speech has occurred. A variation of this algorithm is to use the tongue/jaw motion to make a decision to start the speech recognition (rather than using acoustics as in Fig. III-4 above, and then wait up to 0.5 seconds (or the correct statistical delay) before a second vote is taken using the presence of vocal cord motion or acoustic sound presence to validate the fact that speech has started. Similarly, one decides on end-of-speech, by conducting a final check for non-voiced frictive endings (e.g plural "s" ). If the vote from the unit 8 in Fig. III-4 above shows that speech has started then the conventional acoustic recognizer can start processing the 0.5 seconds of speech recorded in the short term memory. If the vote is that speech has ended, the algorithm stops recognizing speech until a new start condition occurs.

FIG. III-6: Decision Tree for one Non Acoustic Speech Recognition (NASR) algorithm and one conventional acoustic speech recognition (CASR) algorithm. Decision accuracies shown are for illustration only. Examples of top boxes below are box 5 in Fig. III-4 above for the NASR and box 7 also in Fig. III-4 above for CASR. This decision tree is easily extendable to multiple NASRs if several EM sensor are used as shown in the introduction to this document in Fig. I-1.

SCHEMATIC OF VOTING BETWEEN ONE RADAR ALGORITHM AND CONVENTIONAL SPEECH RECOGNITION ALGORITHM, using 97% as decision filter for this example.



III

## B. Background Noise Suppression Algorithm:

There are two issues in background noise suppression: IIIB1: Noise that occurs when the speaker is not speaking but which a CASR confuses with onset or continuing speech, and IIIB2 noise that occurs during the speaker's speech period. In IIIB1, noise occurring while the speaker is not speaking is eliminated as valid speech input by algorithm IIIA1 above.

In case IIIB2 the elimination of acoustic noise (from background) that enters the microphone during speech is more difficult. If a constant high background acoustic level is such that it is comparable to the acoustic input by the speaker into his own microphone, then neither CSRAs nor the simple radar noise rejection algorithms will work. This is equivalent to speaking in a noisy room where your partner can't hear what you are saying. In the case that the exterior noise is short in duration or mostly low in level, it will appear as a short epic of changed signal in the acoustic output of speech processing algorithm illustrated in Fig. III-4 above. The only way to remove this epic or to correct it is to have more accurate data from another sensor, that is trusted so that either a vote can be taken to confirm the acoustic identification or the radar sensor provides such a clear identification that its output is used for the PLU identification. The acquisition of such additional data from radar motion sensors is described below in the multi-position/multi-location algorithms in Section V. However, the algorithm for case IIIB2 is described for use in eliminating short acoustic noise signals by filtering the radar signal to automatically determine if an unusual signal condition happened during the epic that appears unusual to the conventional acoustic speech recognizer.

Three filter algorithms are described below for dealing with a single noise epic:

IIIB2i Radar amplitude change per bin beyond a given threshold (20%),

IIIB2ii Fourier frequency of the dominant amplitude changes by 10%  
(or other user-chosen percent threshold)

IIIB2iii Sudden, unphysical, fitting coefficient change: Up to 20 or more coefficients are used to fit vocal tract models (e.g. LPCs) to fit (and smooth) the digitized radar data in short term memory from each speech epic. For the epic (or epics) during which the noise occurred, the appearance of any change in one or more of the fitting coefficients, greater than that permitted by the normal continuous tract motion model, is labeled as suspect.

If none of the above filters show a discontinuity, then the radar signal was continuous and consistent with ongoing speech, and the recognizer algorithm can vote on the best PLU for the epic based upon the certainty weightings of the radar signal and the acoustic signal or it can reject the acoustic input and then ask the speaker to repeat his or her commands. If the filters show an unusual condition, the radar will confirm the CASR's output

as an indecipherable case, the recognition of the suspect epic will not proceed, and the speaker will be notified

### III-C. Identification of voiced or unvoiced speech-PLC Algorithm:

Vocal folds do not move when non-voiced sounds occur. Examples of 8 voiced and non-voiced PLUs pairs, which are confusing to CASRs, are shown below in table III-1. They are confusing because each pair has the same vocal tract formation, but one is voiced (vocal folds vibrate) and the other is sounded by air rushing through vocal tract constrictions such as almost closed lips as "p" is sounded.

TABLE III-1 (from Olive et al., Acoustics of American English Speech p 24)

voiced	unvoiced	voiced	unvoiced
b	p	v	f
d	t	th (as in then)	th (as in thin)
d	k	z	s
j	c (as in chore)	g (as in garage)	sh (as in shore)

If the CASR algorithm used in conjunction with the radar decision algorithm (shown in Fig. III-6 in IIIA. above) is applied to this decision, it will yield a signature for an acoustic sound that will be somewhat ambiguous between the voiced or unvoiced version. That is, its probability of certainty of identification will be confined to either one or the other PLC of the pair, but the certainty as to which one will be low. The voiced-unvoiced algorithm described here, simply directs the radar data filter in Fig. III-4 to note the presence or absence of a radar signal from vocal fold motion and using the algorithm in Fig. III-6 the CASR library will show two PLCs (e.g. "p" and "b"), and it will test the output of the radar filter which will confirm whether it is the voiced or unvoiced PLC in the epic being examined.

### III D. Pitch Determination Algorithm:

The output from the horizontal processor in Figure III-1 above (also sensor 3 in Fig. I-1) provides the fundamental open and close rate of the vocal cords. The experimental data shown above in Fig. III-3 shows that the signal for the word "two", when Fourier transformed, is easily available to a decision algorithm that selects the peak of vibration above 100 Hz and less than 1kHz. Below we show additional vocal fold vibration data for the two vowels "e" and "u". The algorithm for the pitch recognizer uses the speech epic from unit 2 in Fig. III-4 above and performs first a smoothing filter (e.g. Hamming) and then a Fourier transform. The algorithm performs a search for the highest amplitude signal and then chooses the highest amplitude frequency to be the fundamental pitch. The fundamental frequencies of the two examples shown below differ by a factor of 2 which indicates that "e" is a higher pitched sound than the "u" by the male speaker. This instantaneous pitch information is used by conventional speech recognizers to aid in identifying the PLUs, to train a recognizer to use the natural pitch of the speaker, to determine the excitation function in model based recognition systems, and is used in speech synthesizers to properly drive the vocal track transfer functions used to create natural sounding speech.

fig. III-7a

"eee"

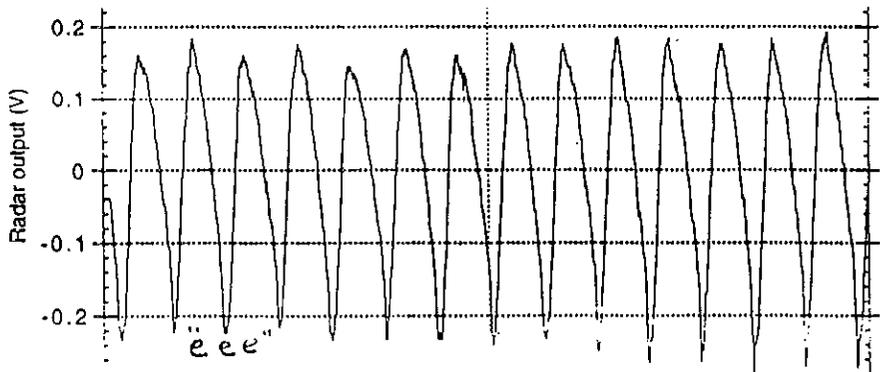
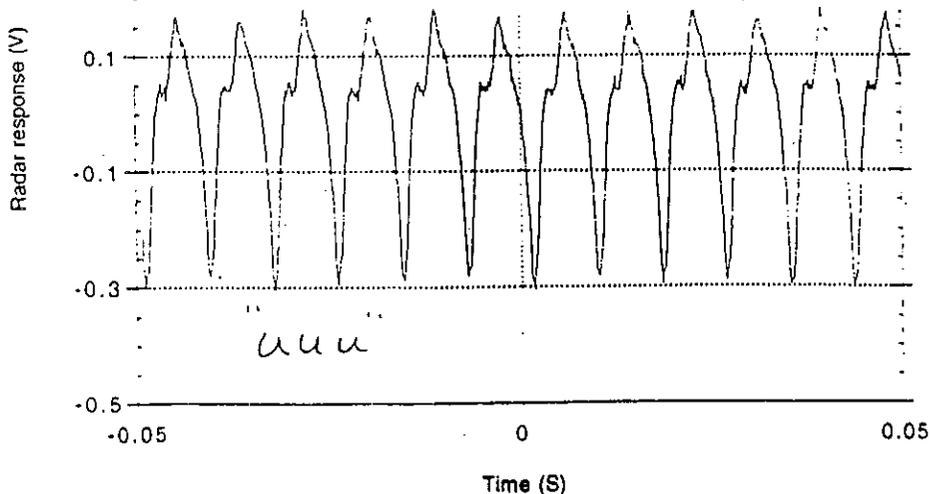


Fig. III-7b.

"uuu"



### III-D. Rate of Speech Indicator Algorithm:

The rate of speech is important for conventional acoustic speech recognizers (CASRs) because they use frequencies derived from time rate of acoustic information flow to identify PLUs. The CASRs use a technique called time warping to align the rates of segments of spoken speech so they can all be recognized with the same recognizers no matter how rapidly or slowly they were spoken. The NASR algorithms described here uses statistics to determine the number of radar recognizable features associated with timing of spoken speech. The general principle is to record the numbers of vocal organ motion events that are uniquely (in a statistical sense) associated with the vocal "flow" of known sounds for the vocabularies being utilized by the user.

#### IIID1: Voiced-Unvoiced Statistics Algorithm:

By measuring the number of times the comparitors numbered 8 and 9 in Fig. III-4 are used in a given segment of speech (e.g. every 4 seconds) one can measure the rate of voiced vs unvoiced PLUs in the 4 second word sets and compare this number against the number in "standard speech". In this algorithm, standard is defined to be the speech rate for which the conventional CASR is set for processing, without its "time-warp" algorithm being used. With this ratio between measured and expected derived from the radar vocal fold sensor, the CASR can be constantly updated with speech rate information.

#### IIID2: Tongue Motion Statistics Algorithm:

By using the tongue motion sensor 2 in Fig. 1 or the tongue motion sensor shown above in Fig. III-1, the rate of tongue motions above a threshold can be measured for each time segment for which speech rate information is needed. This algorithm simply uses a threshold detector in comparator unit 5 in Fig. III-4 above. The number of times the tongue motions exceed the threshold each second is converted to rate of PLUs per second in the speech being spoken. First sections of speech appropriate for the library-vocabulary being used are read into the analyzer, the number of PLUs and the number of threshold triggers are counted for the time interval exercise. The two are compared and a ratio is derived that associates tongue motion threshold evens per second with the rate of PLUs being spoken.

IIID3: Combined Rate Algorithms: It is clear from the two examples directly above that more complex decision trees can be formed by using more than one NASR, each for its own statistical measurements of organ threshold triggering, and then they are combined by statistical averaging to generate a final number for the speech rate signal to the CASR.

### III-E. Rhyming or Difficult Sound Identifier Algorithm

Single organ motion detectors can be used to discriminate between naturally spoken rhyming or otherwise difficult word recognition problems because it is usually the case that the differences in "rhyming" PLUs are associated with one organ motion. The example used in the introduction of the words "saline" and "sailing" are rhymes distinguishable by noting that an EM signal reflection from the tongue tip and the tongue back motions are easily differentiable. Such differentiation occurs as different parts of an organ move and thus reflect the signal at different times during the individual sound epics (for each PLU) in each word cycle.

The use of a single detector is especially useful for limited vocabularies used in specialized applications such as trading stocks or bonds, for banking, for catalogue ordering, for airline system reservations, etc. where very high accuracy on limited word sets are important. The decision tree shown in Fig. III-6 above can be used in the following way. The CASR identifies the nearest PLUs from its library; for rhyming sounds, it will find two or more library PLUs that are statistically close to the incoming sound pattern. However, the library contains along with the CASR identifiers, additional NASR identifying parameters. The additional information contained could include whether the PLU is voiced or not, whether it has front or back tongue positions, etc. The NASR recognizer is consulted for its information from the speech epic, and the PLU fitting both the CASR and the NASR (in a statistical sense) is chosen. In the example, "saline" and "sailing" the CASR has trouble with "ine" and "ing". The NASR recognizer for the tongue would recognize a front position if "ine" were spoken whereas if the tongue were back and closed against the palate, the NASR would indicate "ing". Such "paired" libraries can be built up for all rhyming sounds in a given vocabulary, and in fact can be built up for all rhymed sounds in any language. The number of extra information units to be added to the CASR library of PLU words to accommodate the information for NASR comparison is the number of organ positions being measured by each organ sensor, times the number of PLUs. A typical example would be 4 organs (e.g. vocal folds, tongue, jaw, lips) times 2 positions per organ (e.g. on/off, open/close, up/down) for a total of 8 extra pieces of information per PLU. Since present CASRs use 20 to 30 basis numbers in each characterization "vector" (e.g. in each library location) for each PLU, the addition of 8 more per PLU for comparison to NASR is very easy to accommodate. The extra information could in fact be used to reduce the processing time of the CASR to reach a given accuracy because the NASR data is much more accurate for many words than the CASR, thus less statistical processing is necessary.

## IV. SPEECH TRACT MODEL-BASED ALGORITHMS

### IV-A: CONVENTIONAL MODEL BASED ALGORITHMS

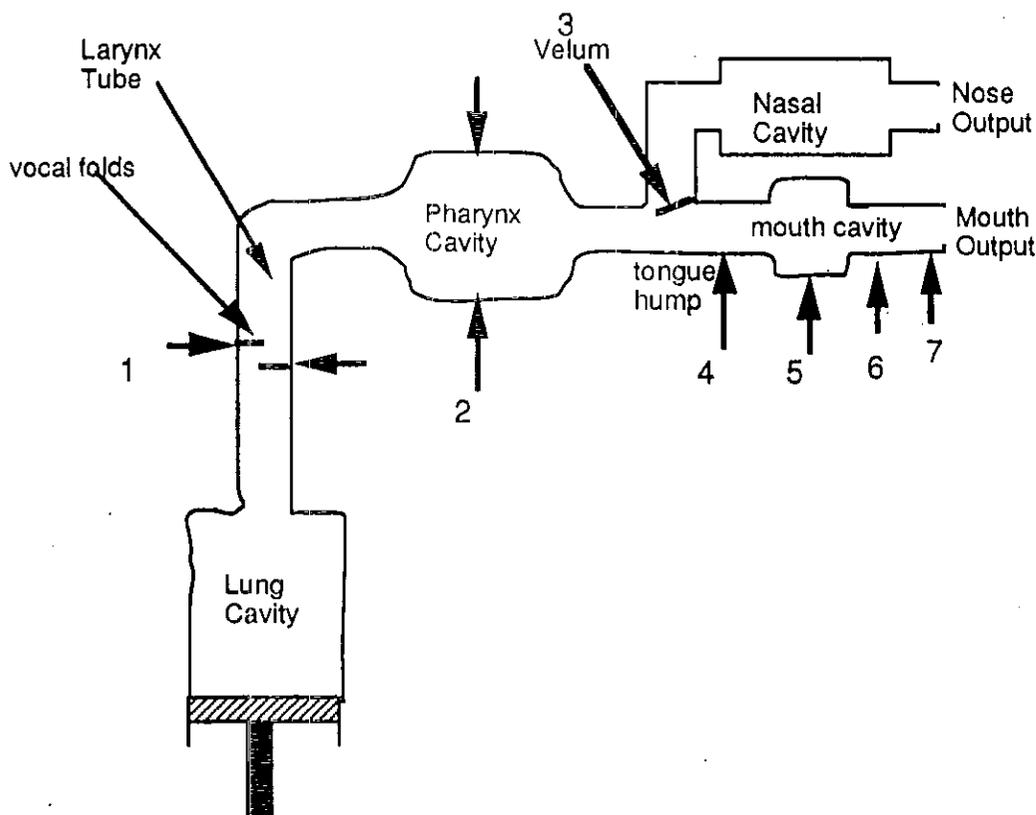
The EM sensors described in the introduction to this document provide data missing from presently used vocal tract models which are in turn used in CASR systems. Presently used CASR systems commonly use LPC (linear predictive coding) models. These models are based on the observation that a given speech epic acoustic output is a linear combination of the outputs from past speech epics. (I use epics to be the time period, typically 10 to 40 milliseconds, during which speech organs change their configuration very slightly, vocal folds excepted) They rely on the fact that the vocal tract organ parameters are slowly varying as the speech organs move to produce each new PLU. The LPC model is a linear polynomial with a series of coefficients ( $a_1$  through  $a_n$ , where  $n =$  or  $> 20$ ) that are fit to the acoustic information from each speech epic. The fitting is done via an analytic, "all-pole" expansion of the cepstral function. The cepstrum is the Fourier transform of the logarithm of the speech power spectrum, which to a certain extent mimics the responsiveness of the human hearing system. See p165, Rabiner and Juang reference 1. The major advantage of this approach to speech recognition is that it is a simple model, it is physical in that the coefficients are constrained to vary slowly as the articulators (i.e. speech organs) change in time, and the coefficients can be easily derived from acoustic spectral data that is obtained during each speech epic. It is also been shown to work well, especially for quasi steady state voiced speech. The reason that models of this nature are desirable for use in real time speech recognition is that they rely on calculating 20 or 30 model based parameters from rather massive amounts of very structured acoustic input information. These coefficients, once calculated, make up the components of a 20 or 30 basis vector characterizing each speech epic. This vector in turn is compared to a library (i.e. or code book) of known vectors for each PLU to be identified, and a statistical speech-PLU identification is made.

The problems with the present model approaches are that the algorithms work backwards from the measured acoustic output to find fundamental parameters which have a high probability of describing the human speech system for each PLU vocalized. The model based algorithm must eliminate the effects of the excitation source (voiced or non voiced), the effects of varying intensities of excitation, and use the remaining information to describe the state of the vocal tract and determine the model (i.e. LPC) coefficients. This approach is necessary when the only information available is acoustic output from the mouth and nose. When additional information is available from EM sensors, a variety of more complex system transfer models can be used. In addition, different models can be used as the vocal tract changes shape to produce quite differently articulated PLUs (sound units). For more detailed descriptions of EM sensors and these models, see Ref. 8.

## IV-B: HUMAN VOCAL SYSTEM

Figure IV-1, below, shows parts of the human vocal system which change in dimension as acoustic speech is formed. It is known from speech studies and from speech synthesis experiments, that knowing the parameters of these human speech organs one can determine the sound being spoken.

Schematic of Vocal Tract  
after Rabiner & Juang Fig. 2.6, p. 17



By knowing the conditions (e.g. dimensions, time variations, etc.) of each of these elements, the human speech sound can be predicted accurately. PLUs can be identified (and, also a very important other application, human speech can be accurately and pleasingly synthesized). EM sensors can determine the conditions of the speech tract with increasing degrees of fidelity as larger numbers of more accurate sensors are employed. This information can be used to augment presently used CASR models such as LPC, it can allow the use of more complex CASR models such as ARMA models (see L. Ng for Auto Regressive Moving Average models, or ARMA, discussions Refs. 8 and 9), it can be used to develop joint NASR/CASR models, or can be used to predict the PLU units without using acoustic speech information at all.

#### IV-C: MODELS:

The essence of a good model is that is easy to use and that it captures the essence of the acoustical physics and the human brain-muscle relationships of speech formation, using as few adjustable parameters as possible.

Furthermore, by "model" one often means a method of describing the relationships between speech organ dimensions or motion conditions and the model parameters describing the speech system model. For example, it will be unphysical for the model to fit a sound with parameters that describe a 4 cm open tongue to palate dimension and a closed jaw. This is accomplished by incorporating mathematical algorithmic constraints on how the model parameters can evolve in time and how they are used as a function of an intended speech unit. The algorithms that work well with EM sensors are those where a small number of model parameters can be constrained by the EM reflection information from a given organ condition. It is not necessary that the EM sensor obtain "picture perfect" dimension information to constrain a model parameter, but that there is a strong correlation between a returned EM signal and the parameter of the model being constrained. An example used below is to use the EM sensors and algorithms described in Section III to measure vocal fold motion, and use this motion information to describe (i.e. define) the excitation function for LPC or ARMA models as to pitch, phase, and amplitude of the driving excitation.

Figure IV-2  
ACOUSTIC TRACT MODELS

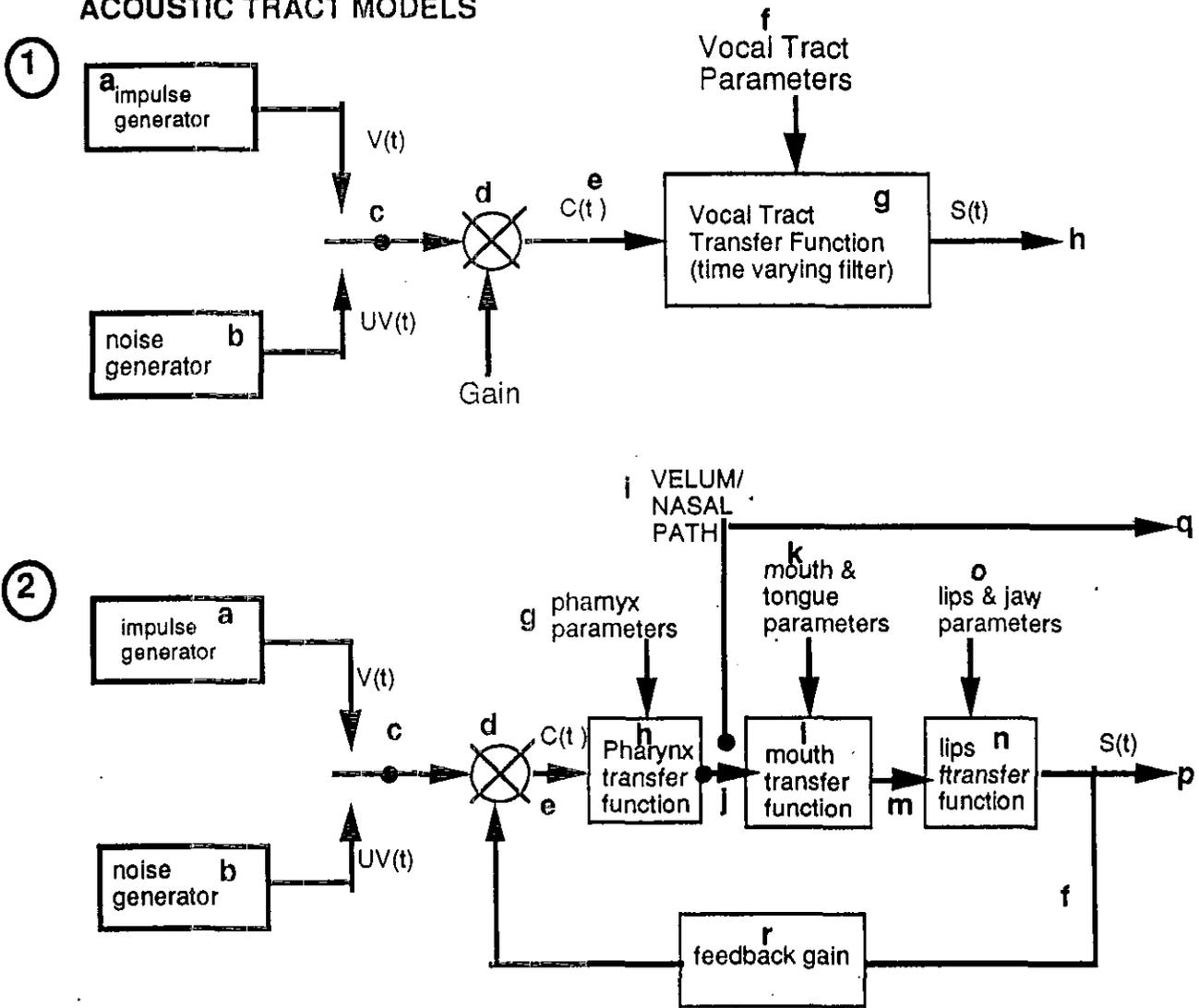


Figure IV-2: Transfer function 1 (top section) is the LPC transfer function where an "all-pole" function is used to modulate the excitation functions to generate speech. The inverse process allows one to determine the coefficients of the LPC function from known speech. The EM sensor information can be used to determine which excitation function is employed by the speaker, and it can be used to constrain parameters in the vocal tract and thus restrict the LPC coefficients more strongly than with the acoustic information alone. Transfer function 2 (bottom section) describes a more complicated transfer function that can include "zeros" and "poles". The letters (i.e. a,b,c,...o) show locations for EM sensor information to be inputted to the transfer function to constrain the model based PLC identification, or to signal the algorithm to change the transfer function.

#### IV-D: USES OF ADDITIONAL EM SENSOR INFORMATION:

The problem with using the simplest of the acoustic tract models is that the information that is desired for use in the recognition procedure ( e.g. the a-coefficients in the LPC model) are in the transfer function, not in the excitation functions, the switches, the feedback, etc. Thus by working backward from the output ( i.e. h in model 1 or p+q in model 2) one has a great deal of excess information that obscures the functions one is trying to determine. NASR systems greatly simplify this problem by allowing the NASR/CASR algorithm user to measure and remove the known functional dependence ( i.e. referring to Fig. IV-2) functions a and/or b in models 1 & 2 and h,i,l,n,f, r in models 2), and thus simplifying the identification of the parameters in the transfer functions used. In reference 8, we show how to obtain the excitation function, characterize it, remove its effects from the acoustic output, and determine aspects of the feedback loop.

The essence of the NASR "vocal tract" algorithmic procedure is to use the additional information obtained by one or more sensors and their associated sub-algorithms to constrain the vocal tract transfer function which best describes the vocal tract's length, width, branching, stop locations, pitch, tension, etc. during the speech epic for which a recognition effort is taking place. Two generalized forms of linear transfer functions, H(z), that can be used in speech recognition are as follows, see Refs. 8 and 9:

Equation IV-2: ARMA (auto regressive moving average) model- ratio of two polynomials. In the z - variable, z denotes delays.

$$H(z) = \frac{b_0 + b_1 z^{**(-1)} + b_2 z^{**(-2)} + b_3 z^{**(-3)} \dots + \dots b_m z^{**(-m)}}{a_0 + a_1 z^{**(-1)} + a_2 z^{**(-2)} + a_3 z^{**(-3)} \dots + \dots a_n z^{**(-n)}}$$

Equation IV-3: Pole-zero transfer function model. Here,  $z = e^{**}(j\omega t)$ , where  $\omega$  is the frequency variable ranges from 0 to  $\pi$ ;  $b_1 \dots b_m$  are the zeros, and  $a_1 \dots a_n$  are the poles.

$$H(z) = \frac{(z-b_1)(z-b_2)(z-b_3) \dots (z-b_m)}{(z-a_1)(z-a_2)(z-a_3) \dots (z-a_n)}$$

Efficient recognition will occur when a unique set of a's and b's for the transform used are identified and associated with a known library (i.e. code-book) value that identify the PLU being articulated. The reason that the

"transform-used" may change is that the EM sensor (used in a single organ mode, see Section III) can determine if the vocal tract is strongly modified. Examples are opening and closing of the velum, opening and closing of the lips, closing of the glottis, etc., all of which dramatically change the length, the boundary conditions, and the branching ratios of the vocal tract model. See Fig. IV-1 above. Trying to fit all of these configurations with one set of parameters, as is now done with CASRs, leads to much poorer speech recognition than desired; it also takes much more computing than desired.

#### IV-E: NASR VOCAL TRACT ALGORITHMIC PROCEDURES

The EM sensor system vocal tract algorithmic procedures are as follows:

IV-E1: Using the algorithms described in the single-organ algorithms in Section-III (e.g. onset-of-speech, voiced-unvoiced detection, and pitch) do:

- a) Choose the excitation function and its pitch for the speech epic being fitted, i.e. choose **a** or **b** based on EM sensor data. If **a** is the operative function, measure the pitch and amplitude.
- b) Measure the excitation functions **a** once (**b** is known to be white noise), model its functional form, take its transform, and divide it into the transform of the measured output,  $S(t)$ . Use real-time measured pitch (Section III-D) or use an estimated pitch to set the excitation frequency values before the division.
- c) Measure the acoustically driven motion of the vocal folds because of feedback. Using single organ algorithms in Section III and appropriate frequencies which are imposed upon the vocal cords by the acoustic pressure waves in the vocal tract resonator. Use this information together with models of the vocal tract to measure the feedback level  $f$  in model 2 in Figure IV-2 above.

IV-E2: Using input from the algorithms in Section III, determine which transfer function is appropriate for the speech epic being analyzed. For example, if the EM sensor/algorithm shows that the velum is open (e.g. a nasal such as "n" is being spoken) and that the tongue tip is up behind the front teeth, this means that the mouth cavity is closed and the nasal cavity is open. This calls for a different transfer function than one which would be used with a completely open, single tract system voicing "aah". (In principle, one or the other of the generalized transfer functions shown above in equation IV-1,2 could be used to describe this tract variation and its coefficients could be deduced from the data). This algorithmic step allows one to select the best transfer function from a catalogue of transfer functions optimized for the 4 or 5 important configurations of the vocal tract. It is

interesting to note that if one only uses LPC polynomials (i.e. all "zero" transfer functions) as CASRs use today, the generalized vocal tract can not be accurately modeled. It is clear that pre-knowledge of the large changes of the vocal tract through the EM sensors enables the algorithm fitting routines to more rapidly converge to its characteristic identification parameters--the a's and b's

IV-E3: By using sensors such as those in the numbered locations in Fig. IV-1, determine the detailed organ conditions (e.g. positions, cavity dimensions, velocities, etc.) and which are represented as transfer function parameters, such as  $f$  in model 1, and  $g, k, o$  in model 2. Using input from the algorithms associated with such sensors, which are described in Section III and Section IV, constrain and determine the a's and b's associated with the appropriate transfer function chosen in IVE-2 above.

IV-E4: Form a characterization vector for the speech epic being recognized by using information from IV-E1 through E3. Choose the operative transfer function, the calculated a's and b's, the excitation source functional form together with its pitch and amplitude, the feedback level, and then form a multi component vector. Compare this vector with the library (i.e. code book) of vectors associated with known PLUs or PLU combinations. Select the best NASR indicated PLU or combination of PLUs, and record the probable error of identification.

IV-E5: Use CSAR to do a statistically separate PLU identification. Use the algorithm described in Section III, Figure III-6 to statistically determine the most likely PLU identification of the joint identification and record the associated error.

#### IV-F: SUMMARY:

A series of algorithmic steps are presented that describe how reflected EM radiation from vocal organs can be used to enhance the accuracy, speed, and range of speech recognition. When these procedures are followed and implemented using conventional system-engineering mathematical procedures, enough additional information is available to the speech recognition process that dramatic improvements in speech recognition will occur.

## V. MULTIPLE ORGAN, MULTIPLE POSITION ALGORITHMS

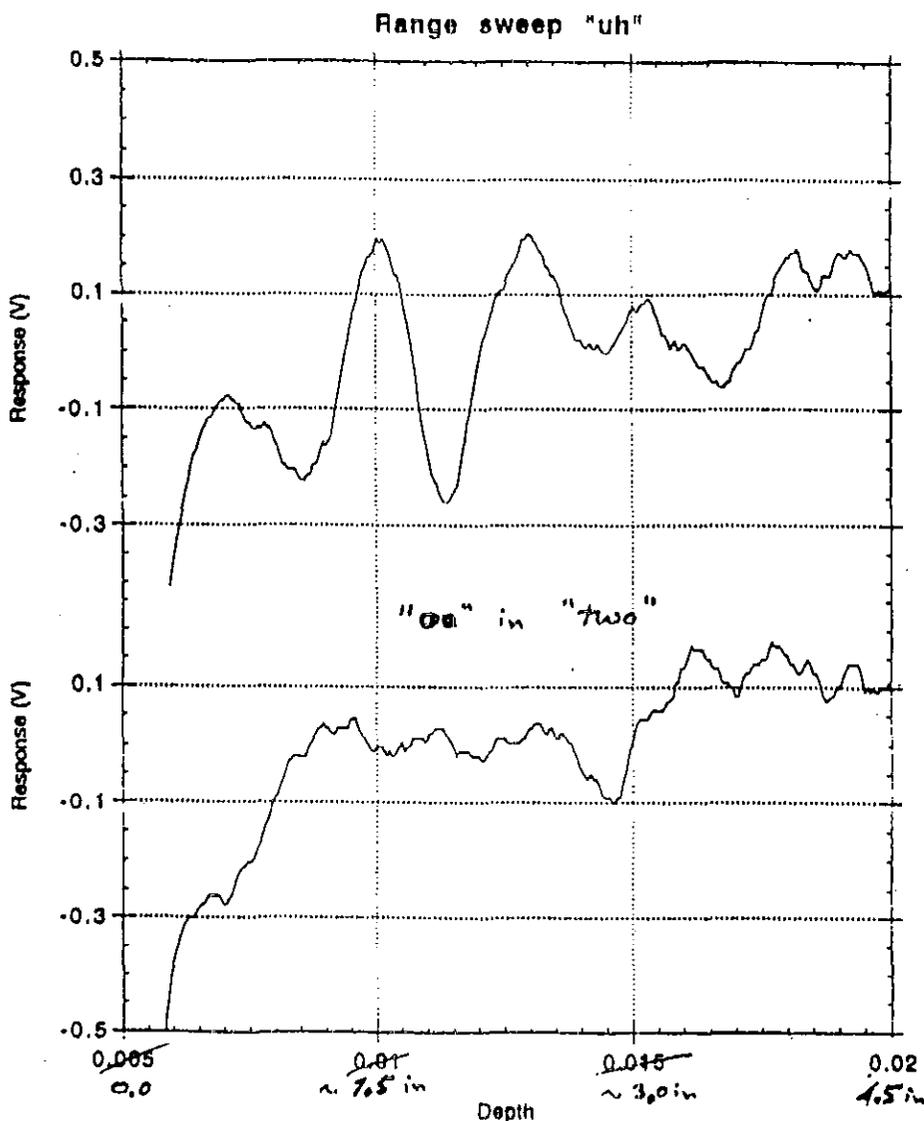
### V-A. MULTIPLE INFORMATION:

Short pulse, high repetition rate ultra-low power EM transmit-receive units have sufficient information gathering capacity that many speech organ interface conditions in the head and neck organs can be gathered within a given speech epic. As discussed above in Section-IIID, reflections from typically 30 locations (or other conditions) along a defined EM wave path can be recorded. As time progresses, and as the vocal organs move to new positions for a new speech epic, the new organ interface conditions (e.g. locations) can be recorded. By recording interface conditions, we mean recording the reflected, processed EM signals associated with the new location configuration. An example of a sensor that is measuring the conditions (in this case distances from the sensor) of several jaw-tongue-palate interfaces during a given epic are shown below in Fig. V-1. However, valuable measurements need not be actual position locations in a photographic sense, but may be complex convolutions of wave resonances, multiple-interface interference effects, whole organ motions, or similar effects. These, less direct data, nevertheless provide information that can uniquely characterize the conditions of the observed organ(s) for the algorithms being used and for the speech recognition market objective being sought. Examples, of very informative, but more complex, convolved EM wave-organ interfaces have been illustrated in Section III (Whole-organ algorithms). There we described how sensors detect gross motions of the vocal folds or tongue motions, which we then showed to provide very valuable information for algorithmic speech recognition decisions. In this section we show how information from multiple sensors, observing several speech organs simultaneously (under some conditions using simultaneously recorded acoustic information) can be statistically "fused" together to generate speech unit (PLU) identification. These procedures also generate other important speech recognition information such as speech start, speech stop, pitch, rate, word separation, organ velocity, speech model coefficients, and others which are very important in assembling the recognized PLUs into words and sentences.

## V-B RANGE GATED MULTIPLE INTERFACE DATA:

An experiment by Holzrichter and McEwan using a range gated radar directed upward into the jaw (Fig. V-1 below) showed a variety of signatures of reflected energy vs time (distance into the head) as function of the dominant sounds from the sounds: "uh" and "uu" as in "two". These signals are associated with differences in jaw up/down, tongue up or down, and the tongue-body (middle tongue) distance to palate (roof of mouth) increasing and decreasing causing changes in resonant reflections. These data clearly show very noticeable changes with different PLUs, which would be dramatically enhanced by subtracting the dotted line noted background.

Fig V-1



### V-C ALGORITHM CONCEPTS:

Given that we can obtain a large number of data from many sensors which are sampling organ interface conditions with a variety of time windows, wavelengths, phases, and processing algorithms, one needs a procedure for integrating the information in such a way that efficient, accurate speech recognition can take place. In this section we show how data taken from demonstrated sensor configurations, together with data from more sophisticated sensor systems which are in the development phase of our work, can be joined together to form multi-component vector identifiers for each speech unit (PLU). This data can be further processed by comparing the data vectors to a library (i.e. code book) of known speech units for the algorithm and sensor suite being used. Together with special weighting factors (built into the library and discussed in Section III and illustrated in Fig. III-6), one can obtain a much higher statistical probability of accurate speech recognition than with present conventional speech recognition systems. In fact, we claim that by using an optimum suite of sensors and algorithms as described in this document, we can provide superior speech recognition to human recognition, <1% error rates.

Below we show a synthesized example of a multi-sensor, multi-interface, multi-organ speech algorithm. It uses a horizontally placed sensor which propagates an EM signal path as in Section I, Fig. I-1 sensor 1. This data is similar to that shown experimentally in Fig V-1, only the concept is expanded to show the potential of the data collection capability of such EM sensor systems when applied to the speech recognition problem. This horizontal, range-gated data set is joined in the algorithm by vocal fold motion data from whole organ sensors (described in Section III). These collective sets of locations (e.g. many organs & many positions) can be correlated with known locations for each phoneme in the set being used (e.g. the "Rabiner set" for English PLUs in Section I, Appendix A), and thus phoneme identification can take place using either electromagnetic information alone, or in conjunction with acoustic speech information. This algorithmic procedure is described in this Section of this document under the category "multiple organ, multiple position" algorithms, and is an extension of the algorithms described in Section III. In addition, the total algorithm eventually used for a working non-acoustic speech recognizer will consist of many sub-algorithm procedures described in the different sections of this document.

## V-D AVAILABLE INFORMATION:

To provide an example of the amount of information available from this multi-organ, multi-sensor technique, consider the available combinations of organ motion and sensor information from very simple technical conditions:

Table V-1

<u>ORGAN</u>	<u>ORGAN CONDITION</u>	<u>INFORMATION UNITS</u>	<u>RELATIVE TIME POSITIONS</u>
vocal folds --		2	
position	open/closed		
rate	high/low	2	
pharynx-glottis	open/ nearly closed	2	2
velum	open/closed	2	2
jaw	up/down	2	2
tongue -- body	up/down	tongue total:	1
tip	up/down	6 data	2
back	up/down		2
lips	open/ nearly closed	2	2

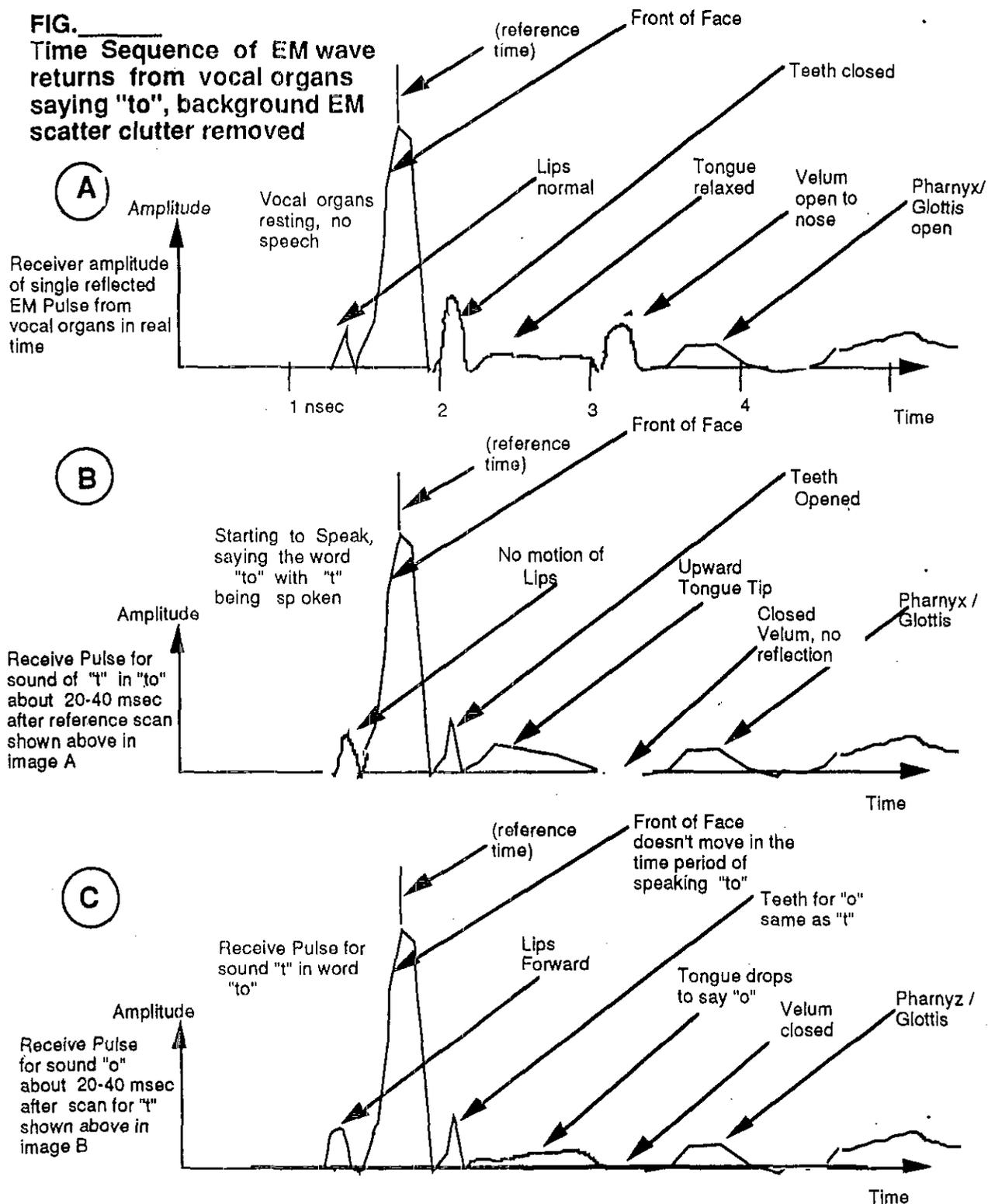
These organ position and velocity conditions all together provide 384 information combinations to be applied to the description of 50 PLU s (speech phonemes) as they are spoken. In addition, the relative time position of each organ condition relative to the other organ conditions within a given speech time-epic also provides information. For a conservative estimate, we note there are at least two important time identification bins per speech epic being considered (1st half and second half). This adds  $2^6 = 64$  more identification units than the 384 estimated above. The total (obtained by multiplying the number of independent parameters together) is 24, 576 potential conditions being described by the relatively simple (i.e.. technically not demanding) set of measurements described above in Table V-1. This number of identifying parameters approaches the number of words used in natural English speech and vastly exceeds the number of speech units normally used for English speech recognition. (e.g. 50 PLU's, 256-512 acoustic units, 2000 English demi-syllables, and 10,000 syllables. See Ref. 1, Rabiner p437.). The complete list of syllables are not used, even though they are the most basic of speech units This is because it is too difficult to identify the measured acoustic identification vector with 10,000 code book vectors. However, it is likely that the NASR measured vector values themselves can be used as addresses to rapidly narrow the search of a corresponding code-book vector in a space as large as 10,000 . PLUs are the present standard set used in speech recognition, thus we use it as the reference set for this document. Research is still

underway to determine the accuracy of determination of the PLUs by the EM sensors-algorithm systems, but our experimental and algorithmic work show that a large number can be uniquely identified. Together, with acoustic information, there is ample data to accurately identify all phonemes and their relationships to the next phoneme in a word or sentence. The objective of applications research will be to identify the minimum set for the market objective, and to minimize the cost, while meeting accuracy, noise rejection, and other conditions.

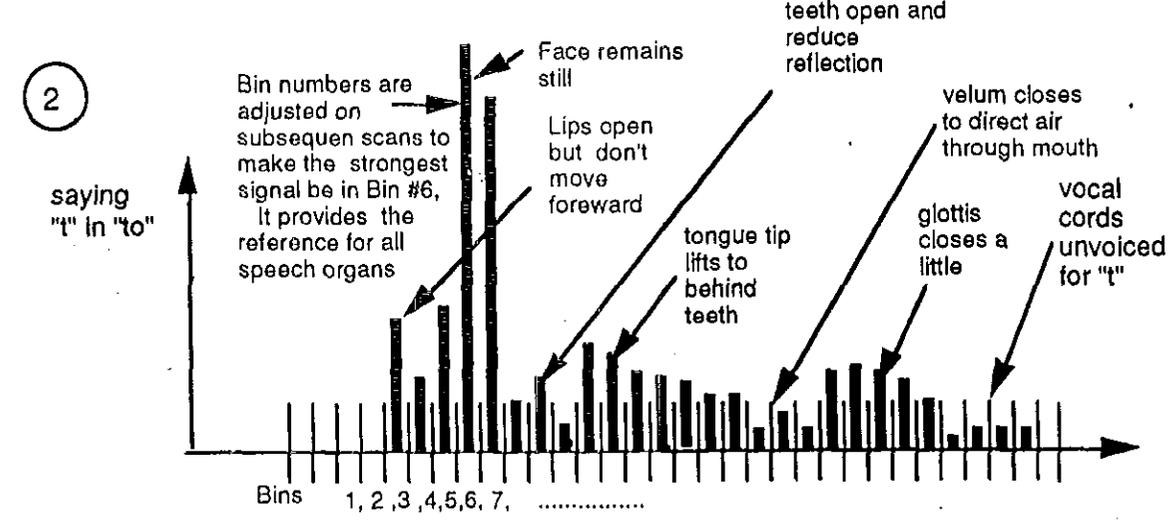
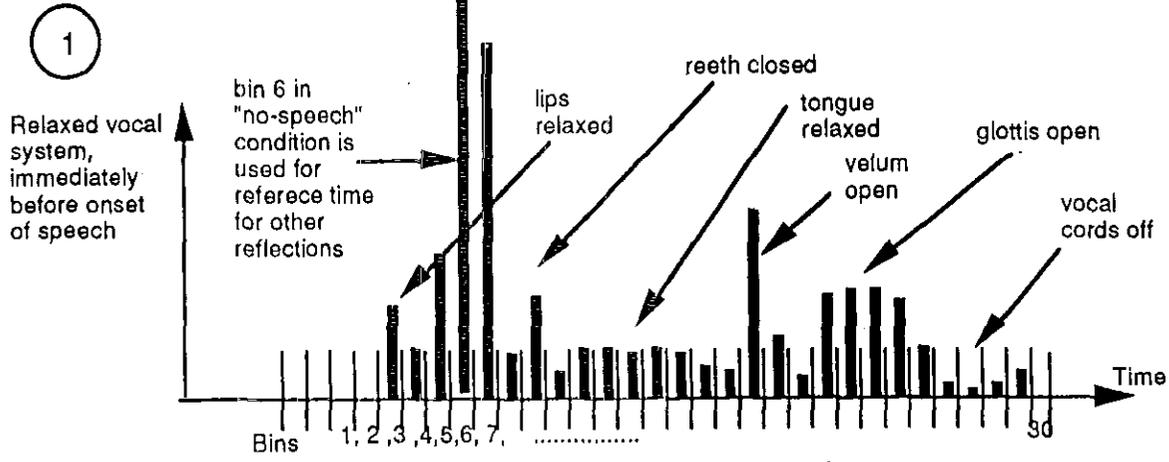
#### V-E ILLUSTRATIVE DATA ACQUISITION AND INFORMATION VECTOR FORMATION:

As an example, by properly choosing a suite of sensors and their wavelengths, as well as pulse format, direction of propagation, receiver conditions such as sample-gate and/or homodyne phase, one can obtain a collection of experimental data as shown above in Fig V-1 and in Section III. The quantization, digitization, averaging, and storing in short and longer term memory of the EM data is described in the following illustrative Figures V-2 through V-4. In particular, these examples show lip-to-throat reflection data vs time (and thus distance) for the spoken word "to", taken (primarily) using a horizontal propagating wave as in Fig. I-1, sensor 1. However, to illustrate the power of multiple-sensor multiple-organ information, vocal cord motion data is added to this set by placing digitized data from its sensor (Fig. I-1, sensor 3) in analog time bins 25 to 28 of the horizontal digital data set illustrated in Fig. V-2 through V-4 below. The last Figure in the series, Fig V-4 shows the strikingly different vector rendition of the data obtained from the PLU "t" vs "o" in the word "to". This data is continuous enough, from bin 1 through 24, that it can provide strong constraints on the model parameters needed for the Vocal Tract Model algorithms described in Section IV. Such models can be used to reduce the number of parameters from 30 in this vector example, to perhaps 10 or 15 truly independent parameters.

**FIG.**  
**Time Sequence of EM wave**  
**returns from vocal organs**  
**saying "to", background EM**  
**scatter clutter removed**

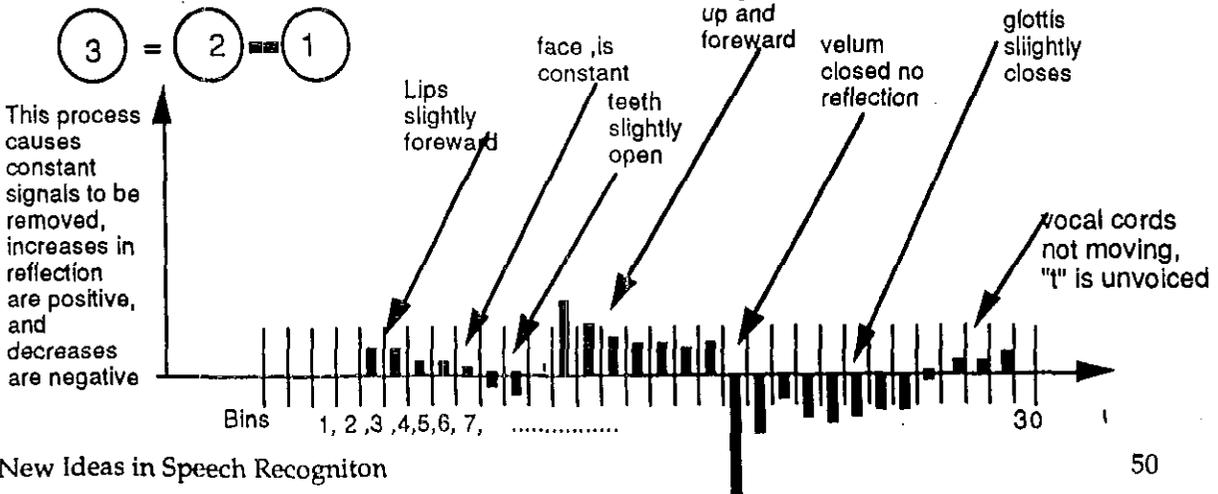


**PROCESSING OF NON ACOUSTIC SPEECH -ORGAN POSITION INFORMATION.**  
 Shows illustrative example of quantized, digitized, averaged, and stored EM received signal information. An algorithm is shown to generate multicomponent vector information for the phonemes "t" and "o" in the two-phoneme word "to". Vocal fold information from a different sensor has been added to bins 26-28

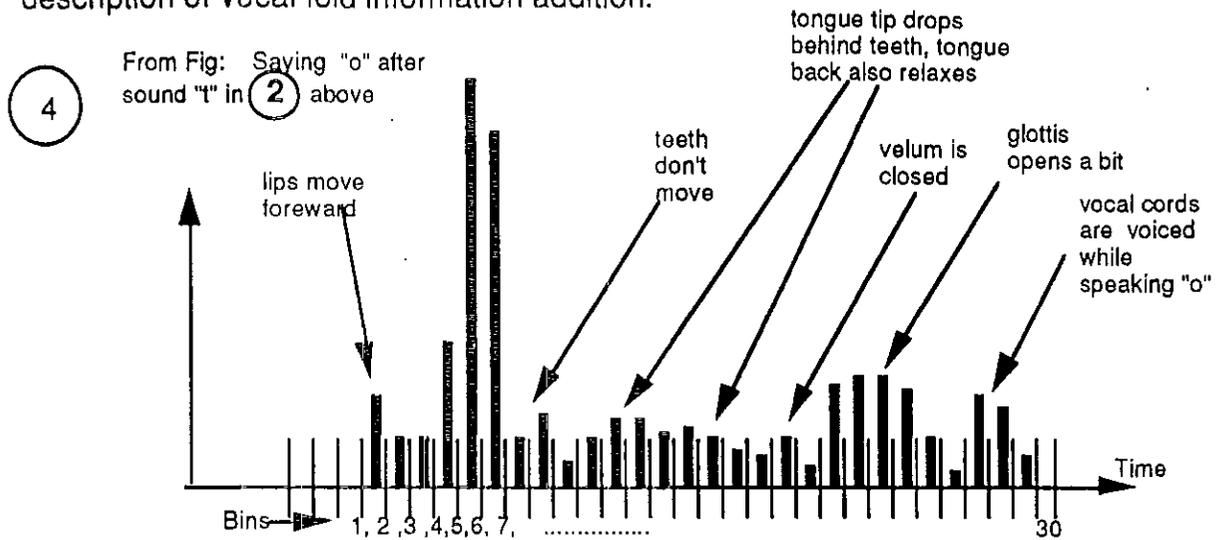


By subtracting the bin numbers in the scans shown in 1 from those in 2 (directly above) one enhances the speech organ position changes and removes static features from the "t".

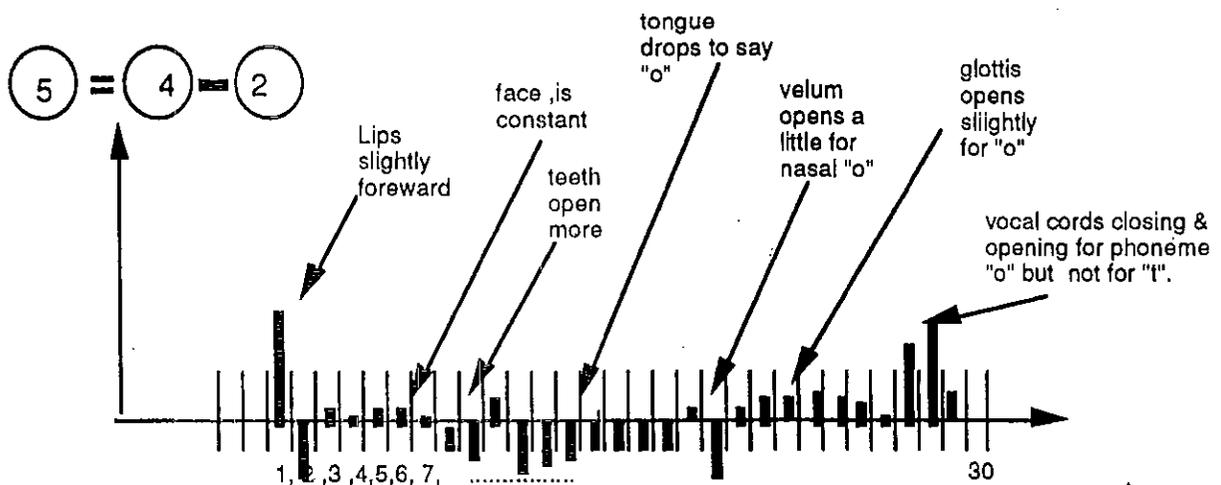
"t" with background removed



**FIG:** \_\_\_\_\_ Formation of PLU vector for "o" in "to", see Fig. above and in in text for description of vocal fold information addition.



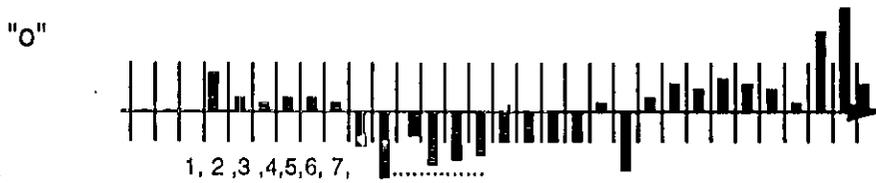
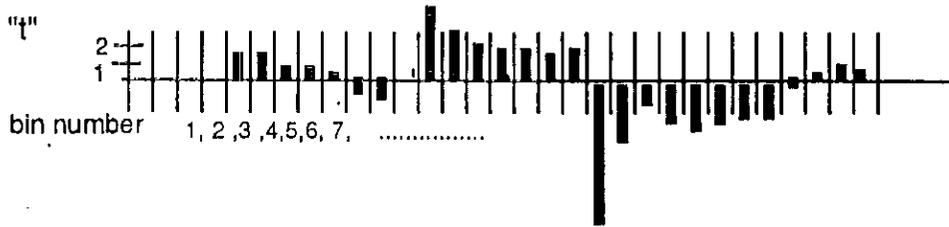
The difference in vocal organ articulator positions from "t" to "o" is shown in 5 below. In this case the scans 4 & 2 would have been taken about 40msec apart. The differences in position, shown by the transition of the tongue, where the signal drops in bin 8 thru bin 17, are very straightforward to detect. Such characteristic patterns as these, can be used to distinguish between the two sounds "t" and "o" in this example.



This array of values vs bin number for "o" forms a vector of 30 components which are quite different than the vector for the sound "t". These can be used to compare to a set of library values of all phoneme values with the given sensor suite and algorithm.

FIG: \_\_\_\_\_ Illustrative Vector patterns for two phonemes "t" and "o" taken using two sensor systems--Horizontal into mouth and Vocal fold on-off, as described in the figures above.

Vector patterns between the phoneme sounds "t" and "o" in the word "to", when compared to a relaxed vocal organ condition. Note large differences in values as they change from positive to negative as the organs deviate in their EM wave scattering strength as they form the articulator positions for each phoneme. If time were included, one would see the "t" being articulated more rapidly than the "o". The actual vector is the series of numerical values in each bin. For example the "t" vector components are, starting with bin 1: (0, 0, 1.5, 1.5, 0.6, 0.6, 0.4, -1.2, -1.5, 0, ..... etc.)



This above example shows how two sensors can be combined to obtain multiple organ information. We show one way to join the two data sets into one set of data to obtain a signature PLU identifier vector (in this case a vector with 30 basis components). This vector would be used to compare to a library of 30 component vectors for known PLUs (i.e. phonemes) to obtain an identification of the best PLU fit. This library would be developed by asking an individual speaker (in the case of a personalized recognition algorithm) to speak a series of words containing all of the PLUs using the above two sensor systems. For an algorithm designed to accommodate a larger number of speakers (for generalized recognition purposes) a cross section of appropriate speakers would be asked to speak all of the phonemes, in appropriate word or sentence situations. The data is then collected, normalized, averaged, and stored in the 30 vector component PLU library. (note: In the library or code book, the 30 component vectors in this example might be joined with more information such as statistical weight, note markers to check other data, etc. for use by the identification algorithm). As shown in section III, Fig. III-6, this data can be statistically compared to simultaneous acoustic data to obtain an improved word identification.

#### V-F LARGER INFORMATION SETS:

The above data manipulation shows how identification vectors for the PLUs can be developed from a given suite of two sensors and two specific sets of formats for EM wave transmission and reception (one range gate, the other homodyne whole organ motion). The above algorithmic description can be extended to a generic multi-organ, multi-position algorithm that uses information from multiple sensors and multiple organs. For example, it can be extended to incorporate several more sensors pointed in other directions, each of which can provide 30 or more bins of data for each speech epic. The additional information can be combined into one vector with 60, 90, or more components. Another approach is to carry several multi-component vectors with each PLU, each vector being especially useful for providing statistically important information for a given set of PLUs and other speech recognition signatures, such as start-of-speech information. In addition, the procedures developed in Section III for combining EM information and acoustic information can be applied directly to the above example where the vector lengths from the sensors are longer (30 vs the 2 or 8 component vector example of a single organ). In particular, in Section III, we showed how NASR data sets (i.e. vectors) can be statistically combined with simultaneous acquired acoustic data sets (CASR) in a fashion described in Section III-C and III-E and in Fig. III-6.

### V-G SUMMARY:

The use of several sensors which are formatted to report on several speech organ conditions and on the relationship of specific organ interfaces with static vocal tract structures provides an abundance of data which has been correlated with PLUs. These information sets when compared to libraries of PLUs and their associated EM wave signatures will make possible acceptable identifications of all PLUs in any language needed. This can be accomplished with or without simultaneous acoustic information, however many of the initial applications are likely to use simultaneously acquired acoustic data. When all available information is used, including acoustic systems (i.e. CASR systems), efficient and low cost speech recognition will be available with almost any degree of desired accuracy. The application and the convenience of use will dictate the system sensor suite, its formats, and algorithmic structure.

## VI. WORD SIGNATURE ALGORITHMS

### VI-A Introduction:

Many series of EM signals can be reflected from the speech organ system during the articulation of a single word. They can be received, averaged, and processed during the time of total word articulation. We often speak at about 2-3 words per second, or 0.500 sec. per word for this example. A word usually consists of 3 to 6 PLUs which we have used in this document as the basic sound unit for speech recognition. The PIUs are connected together in such a way that we recognize them as a meaningful word entity. In an example below, I use 5 PLUs per word. If we use data collection algorithms as described in the Sections above, we described how to collect from 1 to 30 units of information per 15 milliseconds sampling time between speech organ movement. We showed in Section III how whole organ measurements can give 1 unit of information per measurement time and we showed in Section V how multiple organ/multiple interface measurements gave 30 units for each 15 millisecond measurement time. We then described above how to form this EM sensor information into vectors for use in the identification of each PLU; however, we can use the information in a different manner to describe whole words. For example, we may prefer to continuously store the EM sensor signal vectors every 15 milliseconds in a vector location in memory. We then continue to sample, quantize, digitize, average, normalize, and store the following EM sensor information sequentially in the memory every 15 milliseconds, thus generating about 33 vectors per word time of nominally 0.5 sec. These 33 vectors can be used to describe whole words in increasingly complete ways. In the simplest way, we assume the EM sensor information gathered each 15 msec. is represented by one number and we form 33 of these numbers into a 33 component vector for each 0.5 sec spoken word. We could also combine (by averaging or by selecting a key signature or by some other algorithm) the additional information gathered during the 15 millisecond EM sensor acquisition time into one number and store it in the memory location ("bin") to characterized the vocal system condition at a time location during the word sound. We could do many things with all of the data taken each 15 msec, including keeping all information components (e.g. 30 in the multi-sensor, multi-organ example) and arranging these data into 33 columns forming a matrix of information for each moment of word sound. For algorithmic simplicity I describe in the next sub-section an efficient and effective algorithm where single organ sensor data is organized with one number per time time as a word is spoken.

If now we divide the example of 33 information units per word by 5 ( the number of PLUs per word), we note that we are obtaining about 6 identification vectors per PLU speech time unit. These 6 vectors per PLU appear redundant, however when one looks at how PLUs are formed into words, and how words are spoken one after the other there is a need for a

great deal of additional information. Additional information is needed to describe the starting of a new sound (i.e. or PLU), the holding of a sound, the turning off a sound, transitioning to a new sound, and pausing. In the word formation sense, this "extra" information is needed to define word start, word stop, word emphasis, rates of delivery for time "warping" and other important cues for word and sentence construction for the speech recognition process.

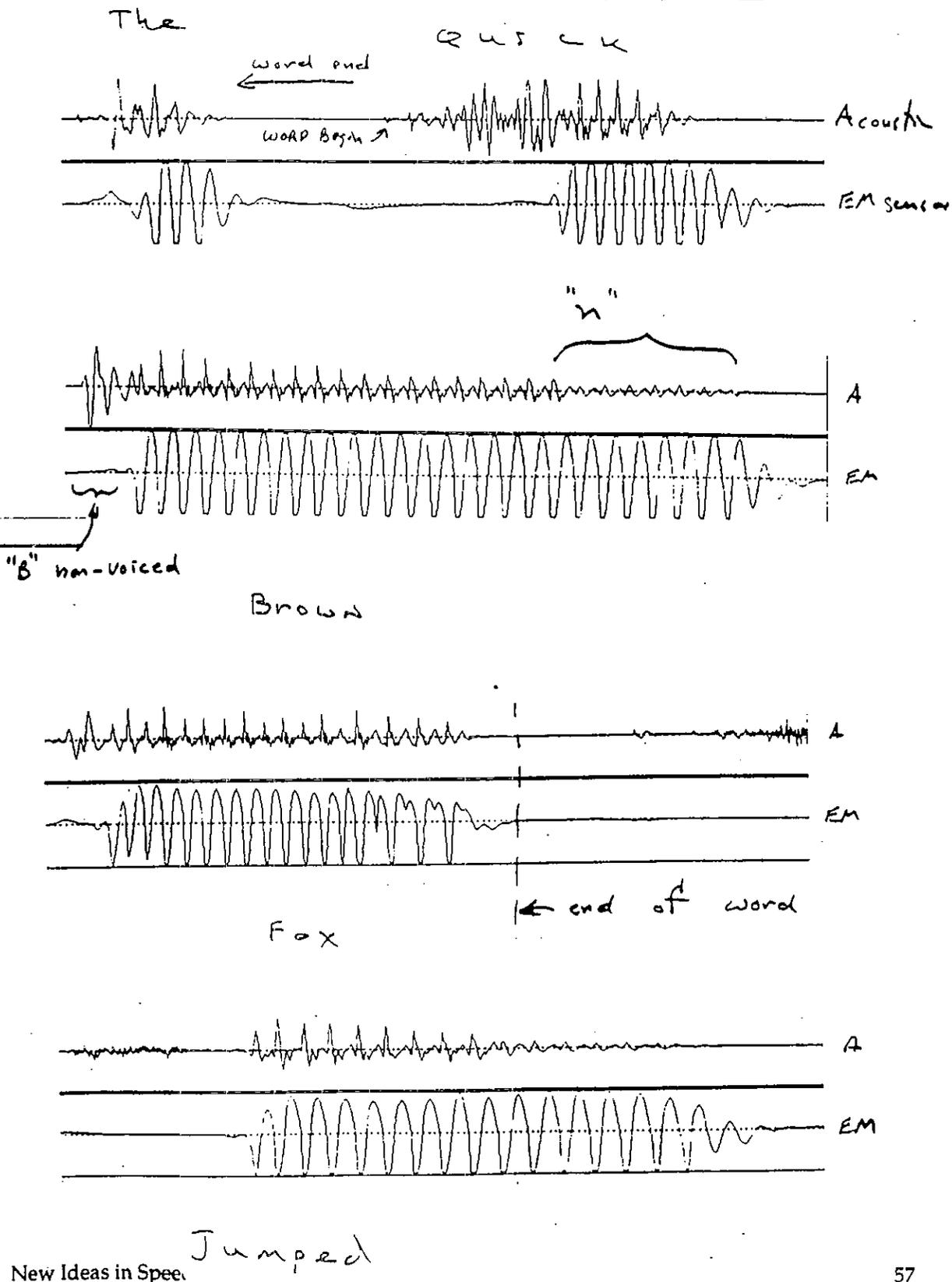
One of the outstanding problems of natural speech is the word-run-together problem. This occurs when there is no break in the acoustic sound between words. We observe this also in the vocal fold sensor information stream when there is no break in vocal fold motion between words (however there is often vocal fold transitions to new pitches, etc.). Only through the use of additional information on articulator motion as provided by the EM sensors and algorithms, can this outstanding problem be solved. The fact is: every change in human speech sound is accompanied by one or more speech organ condition changes. This means that it is now possible for the first time to identify PLU transitions (i.e., all word sound changes), all endings and all new word beginnings using the NASR technology described in this document and referenced documents.

In this document we can not discuss how to assemble PLUs into words; those procedures have been worked out completely for conventional recognition systems and work well once one knows the PLUs and therefore the word starts and stops. They are described well in references on acoustic speech recognition such as the work by Rabiner in Ref. 1 and the references contained therein. However, the added information afforded by the NASR systems discussed above makes possible entirely new types of word identification algorithms which will be utilized by users of NASR systems. Much of the additional information discussed above is illustrated below in Fig V-1 which shows simultaneous acoustic and vocal fold motion data taken by T.McEwan Ref. 10, as a male speaker spoke:

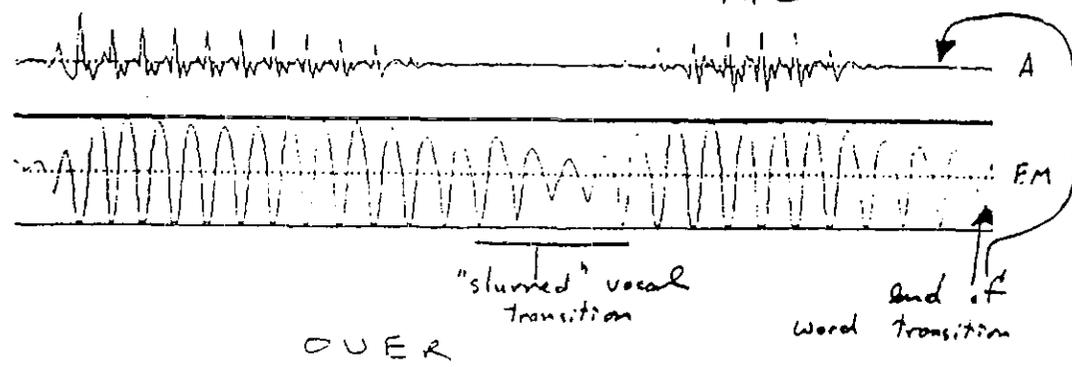
"the quick brown fox jumped over the lazy dog's back"

One sees examples of many of the statements made above, including simultaneous acoustic and vocal fold stops, emphasis changes, PLU breaks, word starts and stops, pre-sound glottal tightening, and vocal fold rate transitions. From other examples given here, and other data we have taken we always see a characteristic EM sensor signal for every sound change that denotes the start or stop of a PLU and therefore the start or stop of a word.

Figure VI-1: Simultaneous acoustic and EM vocal fold motion for the spoken phrase "the quick brown fox jumped over the lazy dog's back":



The

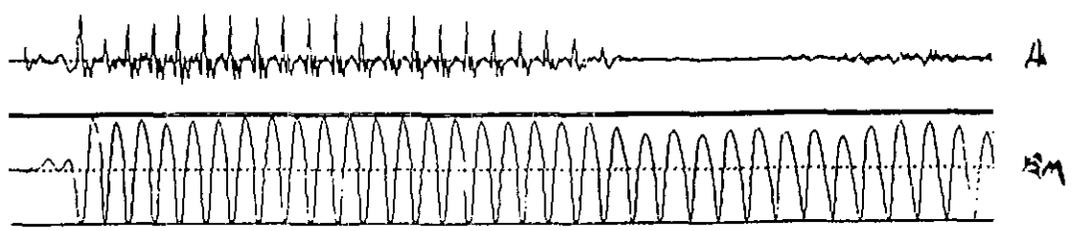


O U E R



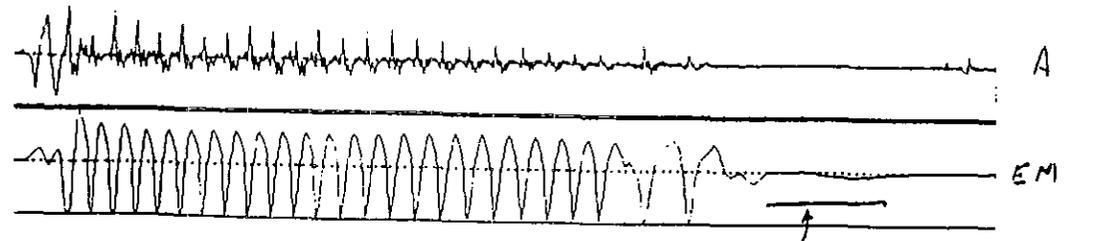
L A Z Y

"ee"



D o g s

"s s s" voiced hard s



w  
B

B A C K

PLU transition

w  
K

## VI-B Word Signatures in Association with Conventional Speech Recognition:

---

There are many applications where very high accuracy recognition of limited vocabularies have great application. Examples are bond trading, airline reservation taking, etc. The vocabularies used in these situations typically have 1000 words or less. Present acoustic processors work on these vocabularies by demanding that the speaker speak clearly, distinctly, and be in a low noise environment. What is needed is additional information that is statistically independent of the acoustic data, so this probability of error can be joined with that of the acoustic information to yield an acceptable quality. Acceptable quality is usually defined to be human speech like which is 1 error in 10,000 words. Such accuracy can be attained by using acoustic speech ( $10^{-1}$  error) and two or more EM sensor/algorithm systems (each with  $3 \times 10^{-2}$  error rates) to obtain a joint recognition error of somewhat less than  $10^{-4}$ .

The algorithm used for the defined vocabulary problem is to take two sets of data, one with CASRs and the other with two or more NCASRs. The word definition and identification is done first by the CASR using an expanded code book which has information in it referring to the expected NASR validation criteria. The NASR data set for each word can contain several types of information. The simplest is that the EM-sensor data is quantized, digitized, averaged over 15 millisecond intervals, and stored in a memory "bin". This continues from the beginning to the end of the word and is used to form a vector of 33 components long (for a 0.5 second maximum duration words). For shorter words, many of the components in the standard vector will be zero, for longer words we use a longer standard vector length. An example of the data that would be quantized, averaged, and stored is shown below in Figure VI-3. In this figure, we show simultaneously sensed acoustic, tongue-jaw position, and vocal fold motion as a speaker says the two words: "sixteen" and "sixty". For this data we would use a vector for a 0.7 sec word length of 50 components, and average the sensor data every 15 milliseconds. The tongue-jaw sensor easily notes the differences between the words. In "sixteen" the word is longer and the tongue-jaw signal stays high longer than in "sixty". Distinguishing between these two words is very important in financial trading. However, these two are often confused with each other by conventional acoustic recognition systems optimized for financial trading, but the words sixty and sixteen are not confused with other words often used in this speech recognition application such as "dollars", "bank", etc. where the CASR does a good job. Note that relatively little extra information is required to "help" the CASR to distinguish between the two acoustically similar sounding words. (In rhyming words or in "difficult words" there is usually only one relatively short information segment to distinguish them from each other, see Rabiner *ibid.* p 291).

Figure VI-3: Distinguishing whole word features for the words "sixteen" and "sixty" as generated by an acoustic sensor, a vocal-fold relative-position vs time EM sensor, and a tongue/jaw position EM sensor.

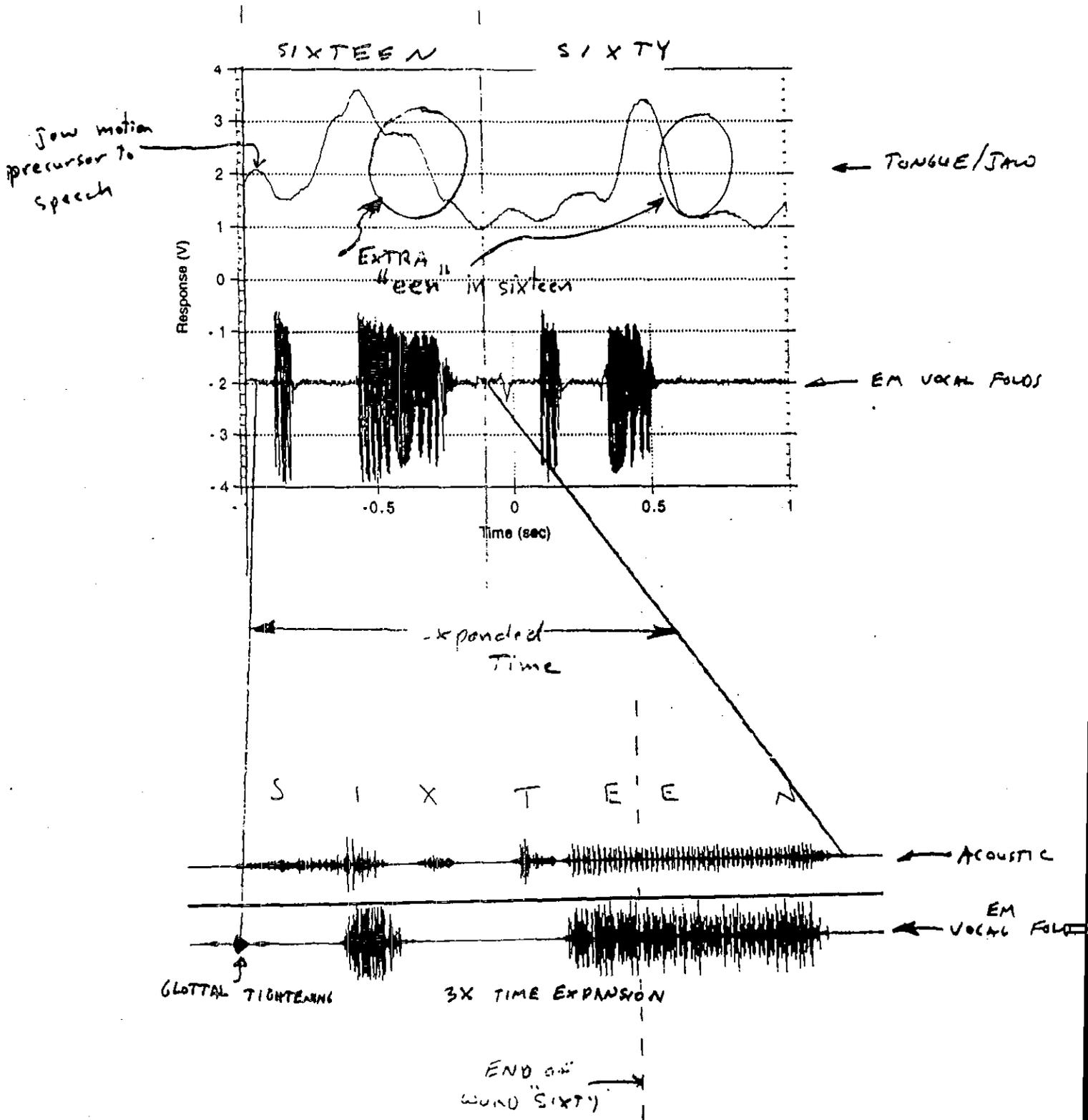


Figure VI-4: A similar set of data which easily shows word distinguishing features is shown below. The two acoustically rhyming words "saline" and "sailing" are sensed with vocal fold and tongue-jaw motion sensors.

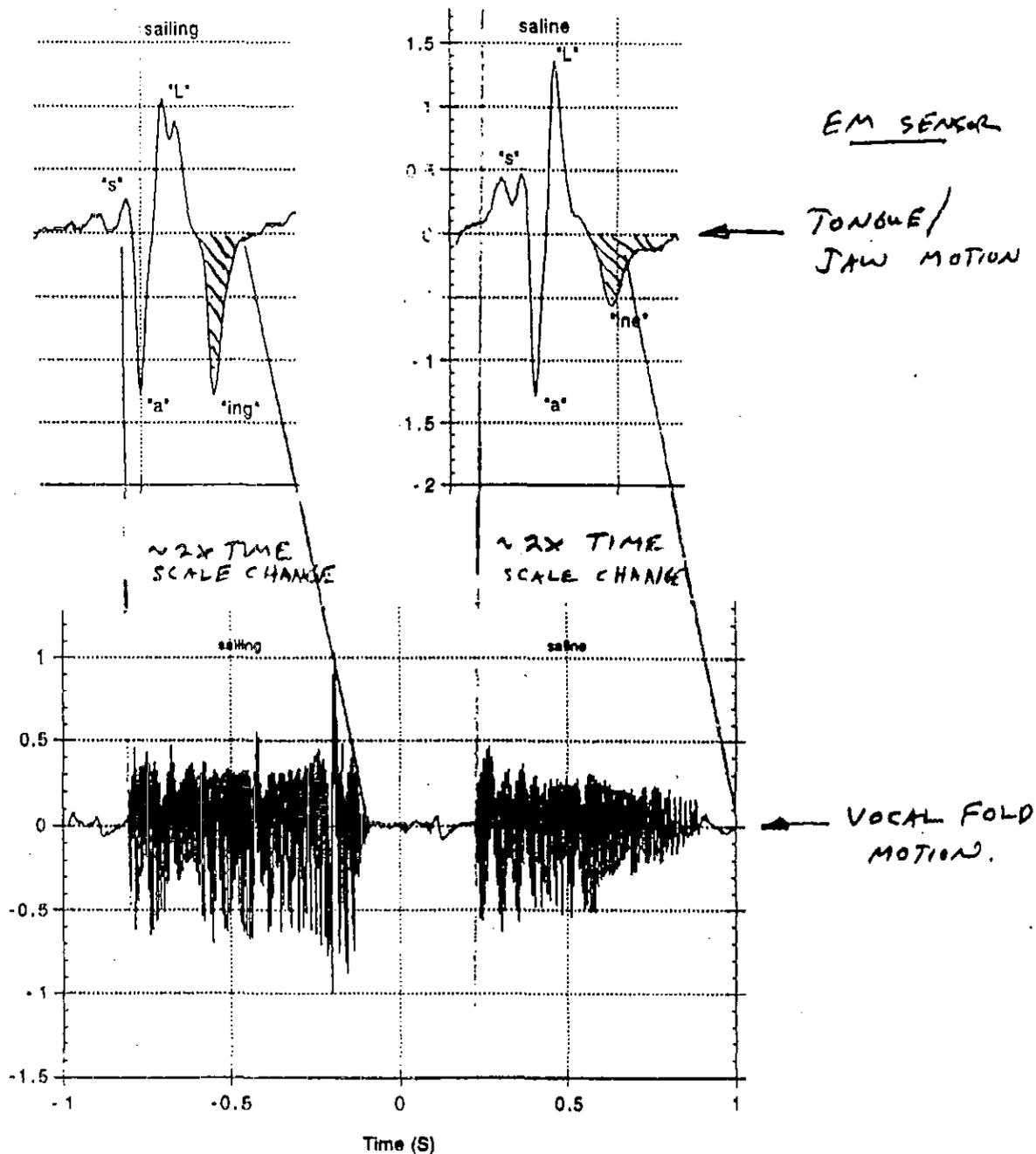
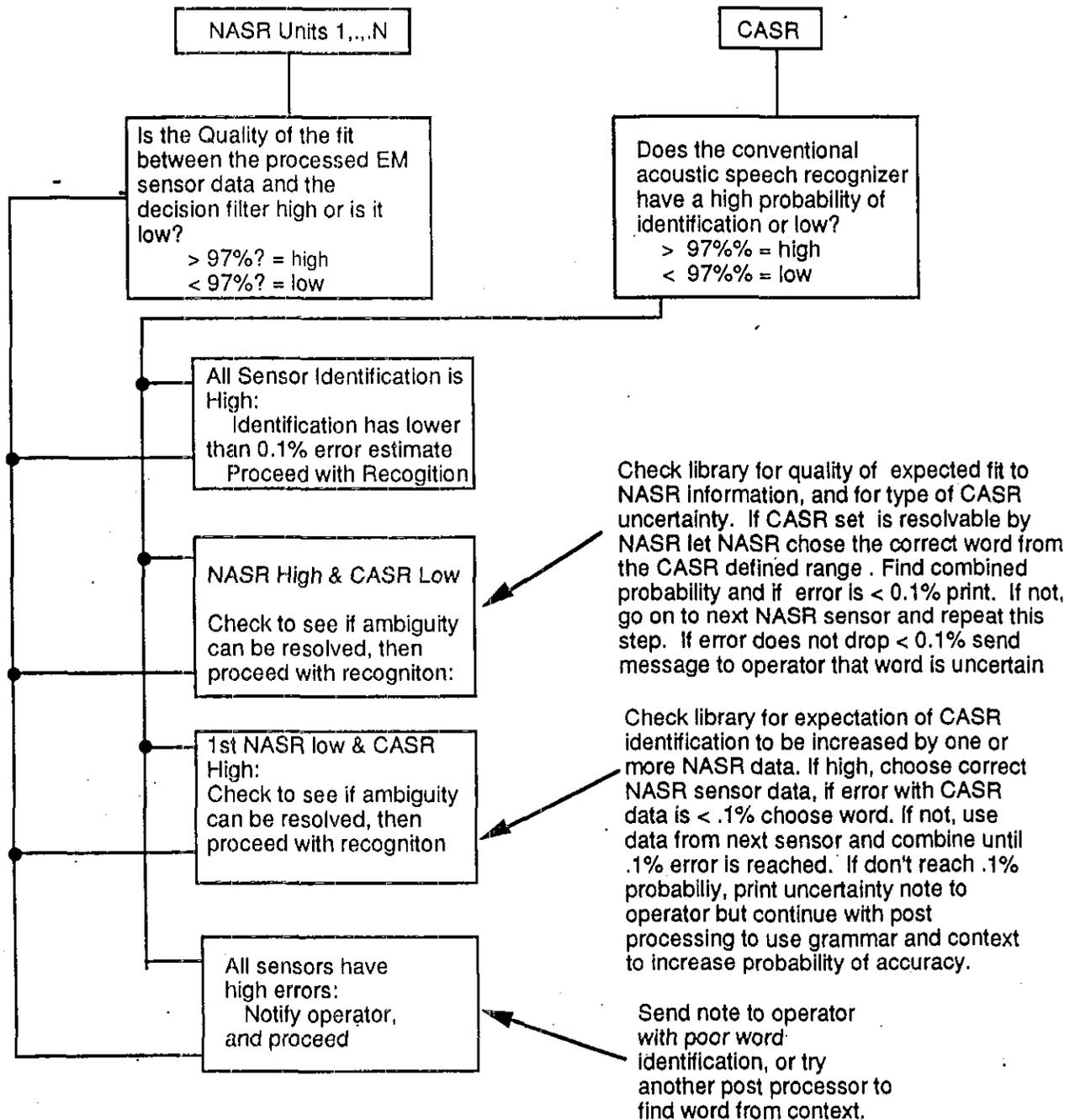


Figure VI-5:

SCHMATIC OF VOTING BETWEEN EM SENSOR-ALGORITHMS AND A CONVENTIONAL SPEECH RECOGNITION ALGORITHM, using 97% as initial decision filter for this example. Two, three, or more EM sensor algorithms outputs can be joined as described. Final joint error probability of 0.1% is illustrative only.



In summary, whole word description vectors can be compared, in a post processor mode, to known vectors for other words in small vocabularies of 100 to 1000 words. That is, after a conventional acoustic speech recognition system makes a decision, the decision is compared against the EM sensor word data to validate the decision. If it is validated with acceptable probabilities, then the word is accepted as recognized, if not, then a best guess is made using the EM-sensor generated data set(s) to discriminate from the subset of words constrained (but not uniquely identified) by the CASR system. Because such acoustically confused words are usually only confused with one or two other "rhyming" words in the constrained set, the EM data easily allows the selection of the correct word with high probability. Above, we show how to use the EM sensor data to resolve the ambiguity in the words "sailing" vs "saline" and "sixteen" vs "sixty". It is straightforward how to extend this algorithmic concept to usefully sized word sets of many hundred to many thousand.

#### VI-C Large Vocabulary, Natural Speech Algorithms:

In natural English speech up to 60,000 words are used when names and technical words are included with standard English. We noted above in section V, that data is available from multiple sensors to distinguish 20,000 different word-sound-units each sampling time period of 15 milliseconds. With the over sampling, and additional information gathering time available during whole word time periods (including pauses between PLUs) several times the 20,000 PLU identifications, up to the 60,000 words needed, are available for word identification. These word identification vectors can be generated by combinations of EM sensor systems and algorithms, and acoustic sensors and their algorithms as described in previous sections. Continued improvements in EM sensors and algorithms, including those to determine word start, rate of speech, pitch, and noise reduction are occurring in our laboratory, and continued reduction in costs of electronic processors and related memories of over 2X per year are occurring in the electronics industry. This leads me to envision that NASR or combined NASR/CASR systems will provide rapid, accurate identification of naturally spoken speech. This will occur via whole word identification algorithms in a process identical to that described in section VI-B above or using PLU, or other sound unit based systems described in Sections III through V above.

## VII. CONCLUSIONS

The above discussed techniques for obtaining speech articulation information and for formatting it such that it can be conveniently associated with stored "code book" information can greatly enhance the accuracy, reduce the computational load, and lower the cost of speech recognition systems. Several algorithmic techniques were described above to do this "recognition". Several more recognition algorithms are being developed and are mentioned here to elicit additional ideas. They include using vectors of information describing pairs of sounds ("diphones") and /or triplets of sounds ("triphones") to use in comparing to the codebook information. These multiple sounds will work especially well with EM sensors because the "silence" phoneme has been a problem in acoustically based systems. However in the EM systems, it is usually associated with a preparatory motion or a relaxing motion of the vocal articulators and is thus useful for the identification of the pair-sound. Similarly, it is known that as a speaker's vocal organs are completing one sound, the non-critical organs (i.e. vocal articulators) are preparing themselves for the next sound. In addition, these motions are sometimes incomplete resulting in poorly articulated sounds. These specific sound pair articulator motions (both complete and incomplete) are only detectable using the EM system described in this report. Another important class of information that becomes available through the EM sensor systems is the articulator "rate of motion" information. While the algorithms described above in the text have mostly used position information in developing the information vectors for subsequent code-book lookup, the rate information as an articulator begins or ends a sound can provide important information that has not been usable. As an example, the "t" sound uses a unique, rapid motion of the tongue tip against the palate behind the teeth. In addition, we have observed that the touch of the tongue to the mouth roof results in a "resonance" as the dielectric tongue tip structure "shorts out" the EM field against the roof of the mouth, and provides a unique reflection of the EM signal. These motion and resonance signatures will provide important information for effective EM speech recognition using new algorithmic techniques.

### Other Applications (see ref. 11 as well)

#### Speech Synthesis

The problems with present speech synthesizer systems are that they have very uninteresting voices (not lifelike), they can not simulate women's speech very well, they are unable to mimic very desirable voices such as those of famous actors, and they have difficulty with complex, run together words. The reason that they have these problems is that they rely either on stored (i.e. recorded) words in a limited vocabulary or on word formation models that are based upon associating acoustic sounds with estimates of the vocal tract mathematical transfer function. Most vocal tract transfer functions are

based upon "all-pole" models, because they are derived from output speech. By comparing with acoustic speech it is not possible to determine transfer function "zeros" and glottal function "zeros". In other words, if no acoustic energy is measured at a given frequency, one doesn't know if it is a consequence of a zero in the transfer function, or if the sound simply wasn't voiced, i.e., a zero in the glottal function. The consequence is that speech synthesizers are not very pleasing in their sound quality. The use of EM sensors described above provides a method of separating the excitation function of the human vocal model system from the transfer function of the human model system, separating them by word units and by neighboring word units, and storing the needed information in realistic data bases for subsequent synthesis.

#### Security:

The information that becomes available through these new channels of EM information provides additional specific information on the idiosyncratic characteristics of each individual speaker. The organ position, velocity, and size information can be recorded in real time along with the acoustic speech information and compared against known information to assist in user (i.e. speaker) identification. The described technology allows one to do the identification with defined word sets or with natural language.

#### Speech Prosthesis:

Many forms of speech defects arise from physical and neurological problems. Deafness makes it difficult for one to speak clearly because feedback is unavailable. A speech organ detection system that identifies the word being attempted and either provides corrective information to the speaker or is used to synthesize speech directly can be of great assistance. Similar applications apply to people with damaged vocal organs, for those who are trying to learn a foreign language, and for those with neurological damage. The EM sensor information can be used to augment or to teach the user to effect improved speech. Feedback can be provided visually, acoustically, tactically, via electrical stimulation, or other techniques.

#### Speech Coding:

The capacity to recognize and synthesize speech in real time, with high accuracy and quality, makes possible many applications in which the spoken speech is "coded" and transmitted as bit sets of vector information units. It is known that the transmission of a written word requires a 100 fold reduction in transmission bandwidth than the spoken word using present voice telephony. Thus the recognition and storage of words, and related speech cues (if needed), can serve an important function in compressed data storage, telephony, and related applications. In particular the ability to measure the human voiced excitation function and thus calculate and characterize the vocal tract transfer function for each speech unit in real time, makes such telephony possible.

### Animal Communications:

The use of the techniques described above should have important applications to animal to animal and animal to human research. Vocal organ motions can be monitored, acoustic communication frequencies can be shifted from one species to another, and other important subjects studied.

### Summary:

I believe that the use of micro power radar vocal organ sensors in conjunction with acoustic sensors, with smaller, more powerful IC processors and memories, and with wireless communication will dramatically change the way we communicate with machines and with each other.

### Appendix A Set of Basic PLUs for American Speech\*

Number	Symbol	Word	Number	Symbol	Word
1	h#	silence	26	k	kick
2	aa	father	27	l	led
3	ae	bat	28	m	mom
4	ah	butt	29	n	no
5	ao	bought	30	ng	sing
6	aw	bough	31	ow	boat
7	ax	again	32	oy	boy
8	axr	diner	33	p	pop
9	ay	bite	34	r	red
10	b	bob	35	s	sis
11	ch	church	36	sh	shoe
12	d	dad	37	t	tot
13	dh	they	38	th	thief
14	eh	bet	49	uh	book
15	el	bottle	40	uw	boot
16	en	button	41	v	very
17	er	bird	42	w	wet
18	ey	bait	43	y	yet
19	f	fief	44	z	zoo
20	g	gag	45	zh	measure
21	hh	hag	46	dx	butter
22	ih	bit	47	nx	center
23	ix	roses			
24	iy	beat			
25	jh	judge			

\*Rabiner & Juang, "Fundamentals of Speech Recognition," p. 438

## REFERENCES

1. "Voice Communications between Humans and Machines" D.B. Roe and J.G. Wilpon, eds. National Academy Press, Washington D.C. 1994
2. L. Rabiner and B.H. Juang "Fundamentals of Speech Recognition" Prentice-Hall, NY, 1993.
3. J.P. Olive, A. Greenwood, & J. Coleman "Acoustics of American English Speech A dynamic approach", Springer, NY, 1993.
4. J.L. Flanagan "Speech Analysis, Synthesis, and Perception", Academic Press Inc. N.Y. also Springer-Verlag, Berlin, 1965,
5. G. Pacun, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data" J. Acoust. Soc. Am. 92, Pt. 1, p. 688, (Aug. 1992)
6. T. McEwan Private communications and "Micro Power Radars" in Popular Science, March, 1995. See also "Ultra-wideband Radar Motion Sensor," US Patent #5,361,070, November 1, 1994
7. M. Skolnik "Radar Handbook" 2 ed, 1990 McGraw-Hill. New York
8. J.F. Holzrichter and L.C. Ng "Speech Synthesis aided by Non-Acoustic Sensors", UCRL- UR-120311 , April 1995
9. L.C. Ng Private Discussions on the use of appropriate models and methods to obtain vocal tract transfer functions by deconvoluting the voiced excitation function from the acoustic speech output.
10. T. McEwan, private communication of simultaneous acoustic and vocal fold motion signals as a male speaker spoke "the quick brown fox jumped over the lazy dog's back"
11. Rabiner, L.R., "Applications of Voice Processing to Telecommunications," IEEE Proceedings, 82(2), 199-228, February 1994