# Ultrawideband Speech Sensing

Ahmed M. Eid and Jon W. Wallace, *Member, IEEE*

***Abstract*—Speech sensing is presented as a novel application of ultrawideband (UWB) technology, possibly improving synthetic speech and speech pathology and allowing silent speech recognition. A finite-difference time-domain (FDTD) model for the human vocal tract is presented and compared to monostatic measurements, indicating that vocal tract dynamics involving the lips and tongue can be identified. A proof-of-concept speech recognition experiment is also presented, showing that even with a very simple algorithm, high match rate is achievable.**

***Index Terms*—Biomedical, near-field propagation, ultrawideband (UWB), UWB antenna.**

## I. Introduction

THIS letter introduces a novel application of ultrawideband (UWB) [1], whose aim is to sense and track the process of human speech, requiring information about the position and movement of lips, tongue, glottis, etc., potentially improving synthetic speech production, speech pathology and therapy, and speech recognition. The method may allow for completely silent voice recognition and silent two-way communications, of interest for law-enforcement or military applications.

Although this present application is related to existing biomedical UWB radar applications [2], there are important differences. In UWB heart-rate and breathing monitoring [3], due to the strongly periodic nature of the signal of interest, only one or two Doppler components must be identified and tracked. In speech sensing, utterances generate nonperiodic signatures that must be identified, where likely 10 to 100 s of parameters in the UWB response should be exploited, and it is unclear whether Doppler-domain or time-domain interpretation is more advantageous.

The purpose of this letter is to introduce the concept of UWB speech sensing and show its potential, but the larger scope of this effort has ties to the extensive progress already made in audio speech analysis and automatic speech recognition (ASR) [4]. Whereas existing audio ASR algorithms typically apply general-purpose pattern recognition algorithms, recent work poses the ASR problem in the more structured framework of identifying cues and gestures from which fundamental phonological segments can be identified [5]. Although a structured analysis of UWB speech-sensing data may allow phonological segments to be directly detected, such a treatment is beyond the scope of this initial study.

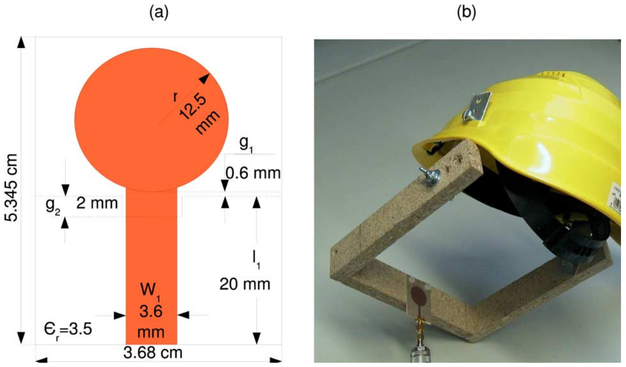

Fig. 1. Broadband monopole: (a) antenna dimensions and (b) antenna headset fixture.

## II. UWB Antenna

Assuming an antenna size on the order of $\lambda/2$ at 3 GHz, it is expected that a compact UWB antenna will have a maximum dimension $L_{\max} \approx \lambda/2 \approx 5$ cm. The beginning of the far-field at a center frequency of 6.5 GHz is $2L_{\max}^2/\lambda \approx 10$ cm, but for optimal sensing, the antenna will likely be much closer, and defining the optimal "pattern" of the antenna in the near-field is difficult. Although optimal antenna design will be the focus of later work, initial investigations employ the simple broadband monopole antenna [6].

Fig. 1(a) depicts the basic geometry of the antenna, consisting of a microstrip line that feeds a disc located past the truncation of the ground plane. The antenna was designed for Rogers 4003C substrate and manually optimized using Agilent ADS to obtain a target reflection below $-10$ dB in the 3- to 10-GHz range, as shown in Fig. 1(a).

Fig. 2 plots the simulated input reflection coefficient $S_{11}$ of the antenna, performed with Agilent ADS (infinite substrate) and finite-difference time-domain (FDTD) (finite substrate), indicating that the target of reflection below $-10$ dB was met. The figure also shows $S_{11}$ of the fabricated antenna measured with a network analyzer in an indoor laboratory environment. Error due to multipath was partially removed by averaging over 150 snapshots of $S_{11}$ taken over a 30-s acquisition (20 ms per sweep), where the antenna was slowly rotated in both azimuth and elevation during this time. Variance of $|S_{11}|$ over the 150 snapshots for all three cases is below $-27$ dB, meaning that averaging mainly helps for nulls at or below this level, but reasonable agreement in the shape of $S_{11}$ is observed.

The other two curves in Fig. 2 show the reflection when the antenna is mounted in a prototype headset (described later), with or without a human subject wearing it, where again a measurement with slow rotation of 30 s was performed. The result indicates that the headset and subject do not significantly reduce the efficiency of the antenna. The pattern of the disc monopole was also simulated in ADS, indicating vertical polarization with a

Manuscript received December 04, 2009. First published January 15, 2010; current version published January 26, 2010.

The authors are with the School of Engineering and Science, Jacobs University, 28759 Bremen, Germany (e-mail: a.eid@jacobs-university.de; wall@ieee.org).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LAWP.2010.2040574

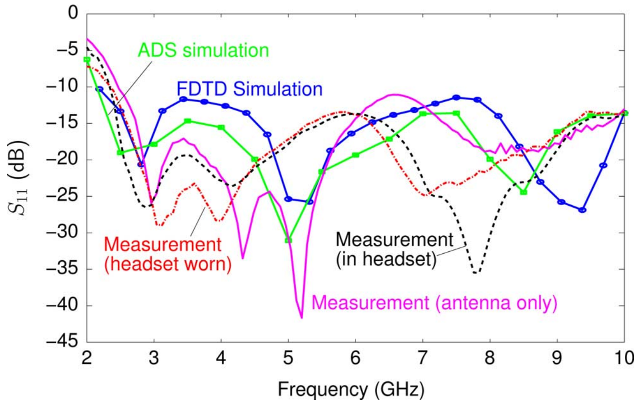

Fig. 2. Simulated and measured reflection coefficient of the broadband monopole.

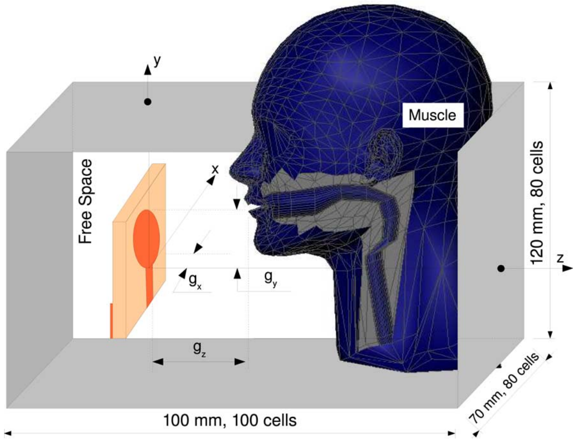

Fig. 3. Head model for FDTD simulations of the human vocal tract.

nearly omnidirectional radiation pattern. Although a directional pattern is arguably more appropriate, the omnidirectional pattern is sufficient for the goals of this study, where a small, lightweight antenna is desirable.

## III. Vocal Tract Modeling and Measurements

In this work, sensing of human speech production is accomplished by placing a single UWB antenna within 1–2 cm of the human mouth, allowing radiated UWB signals to couple with and sense changes in the lips, mouth, tongue, glottis, etc. In this section, a detailed FDTD model of the head and vocal tract is presented, showing that changes in the vocal tract can be sensed from time-varying measurements of the antenna reflection coefficient $\Gamma$. Good agreement between the model and measurements suggests that the measured time-varying response is mainly due to the features of interest and not the surrounding environment.

Fig. 3 depicts the FDTD model that was used in this work, where a human head was generated using a three-dimensional (3D) mesh taken from the MakeHuman project [7] and subsequently edited with Blender 3D computer animation software. Although the MakeHuman head already has a posable jaw and tongue, we were required to create a model of the internal vocal tract in Blender and connect it to the mesh. The vocal tract was modeled using MRI movies [8] for an open and closed mouth.

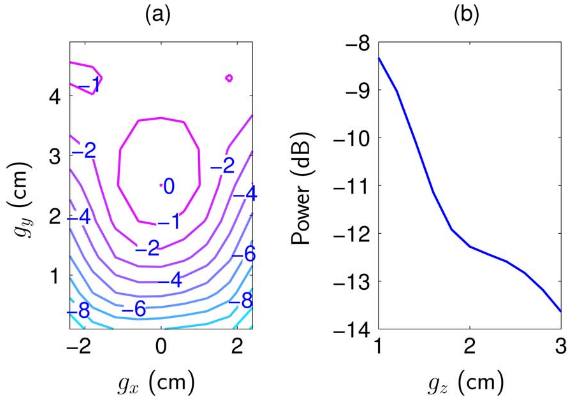

Fig. 4. Relative power level (dB) 2 cm into the mouth for (a) different $g_x, g_y$ and (b) $g_z$ offsets.

As explained in subsequent delta response simulations, the frequency-selective nature of the medium has very little effect on the UWB delta response for this application, and therefore the tissue was modeled using values at 6 GHz ($\epsilon_r = 48.2, \sigma = 5.2$ S/m) taken from [9].

The antenna is placed at a gap of $g_z$ in front of the mouth at horizontal and vertical offsets $g_x$ and $g_y$, respectively, relative to the axis of the mouth. An FDTD simulation domain size of $7 \times 12 \times 10$ cm$^3$ was used, discretized into $80 \times 80 \times 100$ cells in the $x, y$, and $z$ directions, respectively. The PML was 10 cells thick on all sides with a quadratic conductivity gradient and normal reflection coefficient of $10^{-5}$. Note that the parts of the head extending beyond the FDTD simulation domain in Fig. 3 are truncated. The time step was set at 1.6 ps, sufficient for numerical stability. The excitation was a modulated Gaussian pulse with a pulse width (standard deviation) of 80 ps centered at a frequency of 6 GHz, providing adequate coverage of the 3- to 10-GHz range.

### A. *Optimization of Antenna Placement*

The first purpose of the model is to optimize placement of the antenna for peak coupling with the vocal tract. The goal we chose was to maximize the squared magnitude of the electric field intensity integrated over frequency and spatially over a cross section of mouth aperture at a distance of 2 cm into the mouth.

*1) Vertical/Horizontal Offset $g_x, g_y$:* There were 81 FDTD simulations carried out for $g_y \in \{0.1, 0.7, 1.3, \ldots, 4.9\}$ cm and $g_x \in \{-2.4, -1.8, -1.2, \ldots, 2.4\}$ cm with $g_z = 2$ cm, and contours of integrated power at 2 cm into the mouth are plotted in Fig. 4(a), indicating that the best location of the antenna relative to the head is $g_y = 2.5$ cm $= 2r$ and $g_x = 0$ cm.

*2) Antenna-Head Gap $g_z$:* Fig. 4(b) shows the effect of distance between the antenna and the human face for $g_z \in \{1, 1.2, 1.4, \ldots, 3\}$ cm with $g_x = 0$ and $g_y = 2$r. Although having the antenna close is optimal, reasonable coupling and freedom of movement were achieved by using 2 cm separation.

### B. *Vocal Tract Measurements*

In all measurements, the reflection coefficient $\Gamma$ was measured by connecting the UWB sensor to Port 1 of a Rohde & Schwarz VNB20 vector network analyzer, with a frequency

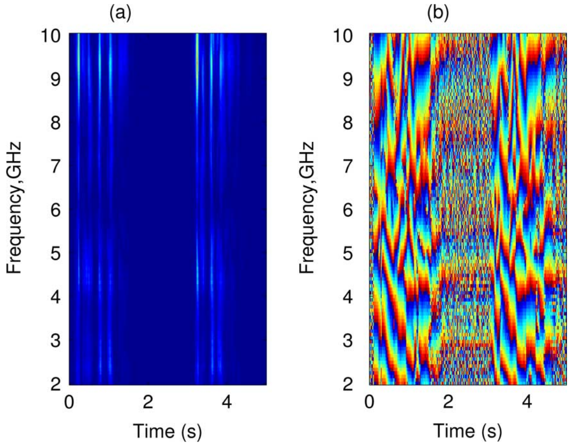

Fig. 5. Delta response of two repeated vocalizations of the word "five." (a) Magnitude. (b) Phase.

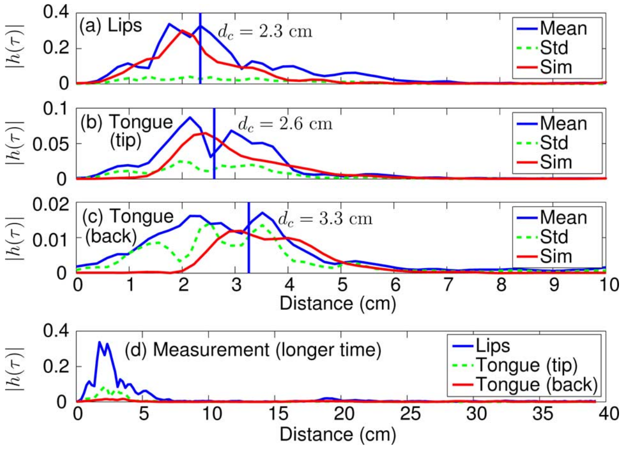

Fig. 6. Time-domain delta response for lips and tongue.

sweep of 101 points over the range of 500 MHz to 10 GHz, a resolution bandwidth of 100 kHz, and transmit power 10 dBm. Note that in subsequent analysis, only the points from 3 to 10 GHz were employed to be compatible with the mask for UWB.

Time variation of the UWB response was captured using an external trigger driven by a digital pattern generator (sync unit) that generated a burst of pulses with a 20-ms period spanning an arbitrary acquisition time. Calibration was performed to remove the effect of the cable up to the UWB sensor. The response was stored on a PC, where $\Gamma_{kn}$ refers to the complex reflection coefficient at frequency index $k$ and time point $n$.

Tracking of speech production is accomplished by a single UWB sensor in the monostatic radar configuration depicted in Fig. 1(b), where UWB signals are transmitted from the antenna; interact with the face, mouth, and throat; and are reflected and measured by the UWB sensor in the reflection coefficient $\Gamma$. In order to concentrate on the time-varying behavior of the vocal channel, the derivative of the response, referred to as *delta response*, was computed as $D_{kn} = \Gamma_{k,n+1} - \Gamma_{kn}$, which removes static effects due to antenna mismatch or environmental scattering.

Fig. 5 depicts an example delta response for two repeated vocalizations of the word "five" spaced by 1 s. During periods of silence at the start and end of the record, minimal change leads to a delta response with low amplitude and random phase. During vocalization, the weakly modulated return signal can be identified, and significant information about the state of the mouth and vocal tract is obtained.

Although identifying the complicated dependence of $\Gamma$ on the state of the vocal tract is a subject of our ongoing work, we present some experiments that were performed to test whether obvious pronounced changes in the vocal tract produce similar responses for both the model and measurements, yielding confidence that the specific target feature is sensed, and not some other aspect of the body or the measurement environment. In each case, the delta response between two vocal states is computed. For example, for closed- to open-mouth transition, $d_k = D_{k1} = \Gamma_k|_{\text{open}} - \Gamma_k|_{\text{closed}}$. Also, each measurement was repeated 20 times, and the mean and standard deviation of the responses are computed, indicating repeatability.

Fig. 6 compares measurements and FDTD simulations in terms of the difference of $\Gamma$ (delta response) for two extreme vocal states. To allow the range (i.e., delay) of features that significantly impact the delta response to be identified, each frequency-domain delta response is transformed to a complex baseband time-domain response $h(\tau)$. For a hypothetical continuous-frequency delta response $d(f)$, the inverse Fourier transform of the equivalent complex baseband signal is $h(\tau) = \int_{-\infty}^{\infty} w(f)d(f)\exp[j2\pi(f - f_0)\tau]\ df$, where $f$ is frequency, $f_0 = 6.5$ GHz is the center frequency, and the window function $w(f)$ is a Hamming window spanning 3 to 10 GHz. For the discrete frequency samples obtained from our measurements, the inverse Fourier transform is approximated in a conventional manner by applying a discrete Hamming window, zero padding, and an inverse fast Fourier transform.

The time axis is transformed to one-way distance by assuming free-space propagation. Note that all simulated curves have been shifted by an identical amount to achieve the best fit for the lip movement case in (a), removing any static differences due to source position. Also, the RMS delay distance $d_c$ for measurements has been computed and is indicated as a vertical line in the plots. Although Fig. 6(a)–(c) look similar (a cluster of energy appears over some range), the observation is that changing incrementally deeper parts of the vocal tract leads to energy in the delta response at a longer delay and a weaker level, strongly suggesting that the features of interest are responsible for the energy we measure in the delta response.

Fig. 6(a) plots the case of lip movement where the two states are open mouth (3 cm diameter) and closed mouth with the tongue resting on the bottom of the mouth. As can be seen, the responses for the measured and simulated vocal tract are very similar, with a slightly higher response for the measurement. Also, the measurement is quite repeatable, evidenced by the low standard deviation.

Fig. 6(b) shows the result for an open mouth (3 cm), where the two states are the tip of the tongue resting on the bottom of the mouth versus touching the roof of the mouth. Both measurement and simulation show a shift of the response to the right relative to case (a) and good agreement in level.

Fig. 6(c) depicts the result for the back of the tongue up or down with a mouth opening of 3 cm diameter. For a lowered tongue, the state of the vocal tract is that of the vowel "ah" and for raised that of the initial g on "gah." Care is taken to keep the tongue and jaw as still as possible for the two states. Both measurement and simulations show significant energy coming from between 3–6 cm, roughly corresponding to the distance to the back of the tongue. Both the unexpected measured energy for distance less than 3 cm, as well as the increased standard deviation, are likely due to the difficulty of keeping the mouth completely still while changing the back of the tongue.

Fig. 6(d) shows the mean delta response for the three cases for a longer observation time, allowing us to assess the impact of more distant environmental features (such as nearby equipment) on the measurement. The plot suggests that environmental scatterers impact the delta response weakly compared with vocal tract changes. Also note that delta responses for cases (a)–(c) were simulated with the material properties of muscle computed at the band edges (3 and 10 GHz), and negligible shift in the delta response curves suggests that material dispersion is of lesser importance in this application.

## IV. UWB Speech-Recognition Experiment

As an initial test of the potential of speech sensing with UWB, a simple speech recognition experiment was conducted. This experiment was performed using actual measured responses only since the vocal tract model is too simple to simulate detailed movement of the tongue and mouth. Although recognition of phonological segments [5] such as vowels and consonants is a goal of this effort, this requires extensive work to identify the main features of interest in UWB speech responses and how they are connected to speech processes.

Here, we consider only simple whole-word recognition by template matching, where a dictionary is formed by recording the UWB response of several words spoken many times each, where $D_{kn}^{(m,p)}$ is the delta response of the $m$th word and $p$th trial. Next, the UWB response of a new vocalization is recorded, denoted $D_{kn}$, and compared to the dictionary. Recognition is accomplished by minimizing a simple sliding distance metric, in which the recognized word index is

$$m = \arg\min_{m}\min_{p}\min_{i} \frac{\sum_{k,n}\left|D_{kn} - D_{k,n+i}^{(m,p)}\right|^2}{\sqrt{\left(\sum_{k,n}|D_{kn}|^2\right)\left(\sum_{k,n}\left|D_{kn}^{(m,p)}\right|^2\right)}} \tag{1}$$

where the sums span the ranges $n \in [1, N]$ and $k \in [1, N_F]$, and $N$ and $N_F$ are the number of time and frequency samples, respectively.

Table I shows the results of a speech recognition experiment involving speaking the integers "zero" through "nine," where 30 trials of each word are stored in the dictionary. After recording the complete dictionary, each word on the left column is spoken 25 times, and the number of times it is matched with each word in the dictionary is tabulated. The poorest performance occurs for "six," possibly due to the fact that its vocalization is very short, yielding a rather weak signature. The average rate of recognition mismatch is low (around 7%). This is quite encouraging, given the simplicity of the algorithm.

TABLE I
Example Speech Recognition Experiment

| Word | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 21 | 0 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 1 | 0 | 24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 1 | 0 | 0 | 0 | 24 | 0 | 0 | 0 | 0 | 0 |
| **5** | 0 | 0 | 0 | 0 | 0 | 23 | 1 | 0 | 1 | 0 |
| **6** | 1 | 0 | 0 | 0 | 3 | 0 | 18 | 1 | 1 | 1 |
| **7** | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 22 | 0 | 0 |
| **8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 |
| **9** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 |

The effect of the environment on speech recognition was checked by applying a time-gate filter to the measurements to remove signals arriving after 2 ns, thus excluding objects more than 30 cm away. The effect was that the words "two," "six," and "eight" were recognized correctly for 23, 19, and 24 of the trials (the other words were unaffected), respectively, indicating no improvement in the algorithm and suggesting very little impact of environmental scatterers.

## V. Conclusion

This letter explored the novel idea of UWB speech sensing, with the goal of improving synthetic speech and speech pathology as well as allowing silent speech recognition. A detailed FDTD model was employed, showing good agreement in the delta response as compared with measurement of changes in the mouth and tongue. A proof-of-concept speech recognition experiment showed that speech detection with the proposed method is possible. Future goals of this work include the development of more optimal directional sensors for speech sensing, improved modeling and characterization of the connection between speech and UWB responses, and improved algorithms for UWB speech recognition.

## References

[1] L. Yang and G. B. Giannakis, "Ultra-wideband communications: An idea whose time has come," *IEEE Signal Process. Mag.*, vol. 21, no. 6, pp. 26–54, Nov. 2004.

[2] E. M. Staderini, "UWB radars in medicine," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 17, no. 1, pp. 13–18, Jan. 2002.

[3] F. Thiel, M. Hein, J. Sachs, U. Schwarz, and F. Seifert, "Physiological signatures monitored by ultra-wideband-radar validated by magnetic resonance imaging," in *Proc. 2008 IEEE Int. Conf. Ultra-Wideband*, Sep. 10–12, 2008, pp. 105–108.

[4] D. O'Shaughnessy, "Automatic speech recognition: History, methods and challenges," *Pattern Recogn.*, vol. 41, pp. 2965–2979, Oct. 2008.

[5] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1872–1891, Apr. 2002.

[6] J. X. Liang, C. C. Chian, X. D. Chen, and C. G. Parini, "Study of a printed circular disc monopole antenna for UWB systems," *IEEE Trans. Antennas Propag.*, vol. 53, no. 11, pp. 3500–3504, Nov. 2005.

[7] "Make Human," [Online]. Available: http://www.makehuman.org/

[8] H. Shinagawa, T. Ono, E.-I. Honda, S. Masaki, Y. Shimada, I. Fujimoto, T. Sasaki, A. Iriki, and K. Ohyama, "Dynamic analysis of articulatory movement using magnetic resonance imaging movies: Methods and implications in cleft lip and palate," *Cleft Palate-Craniofacial J.*, vol. 42, pp. 225–230, May 2005.

[9] "Dielectric properties of body tissues," Italian National Research Council, Institute for Applied Physics [Online]. Available: http://niremf.ifac.cnr.it/tissprop