Ultra-wideband Sensing of Speech Production

Ahmed M. Eid and Jon W. Wallace*
Jacobs University Bremen
School of Engineering and Science
Campus Ring 1, 28759 Bremen, Germany
a.mossa@jacobs-university.de, wall@ieee.org

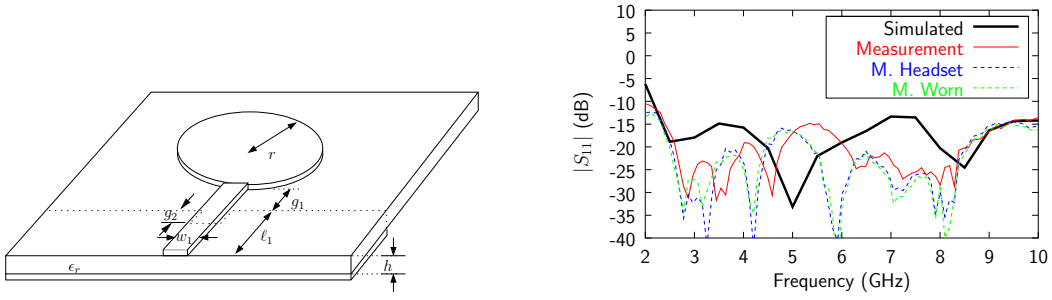
Introduction

Ultra-wideband (UWB) technology has been identified as an attractive solution for certain biomedical sensing applications due to its safety, no requirement of physical contact, real-time capability, and low cost [1]. This work considers UWB sensing of human speech production, potentially allowing significant improvements in synthetic speech production, speech pathology, and speech recognition. Results of a simple speech recognition experiment are presented. A multi-section cylindrical waveguide model is proposed to approximate the human vocal tract for future inverse scatting studies.

UWB Sensor

A multitude of antenna geometries have been proposed and developed for operation in the UWB range from 3 to 10 GHz, such as traditional broadband horn, biconical, and discone antennas. The broadband monopole antenna is used in this work due to its compact planar design and ease of fabrication. Figure 1(a) depicts the chosen broadband monopole antenna, consisting of a microstrip line which feeds a disc located past the truncation of the ground plane. Rogers 4003C high-frequency substrate was used with $\epsilon_r = 3.5$ and $h = 1.524$ mm and a 50 Ω microstrip feed line required $w_1 = 3.6$ mm. Optimization with Agilent ADS to minimize return loss in the range of 3 to 10 GHz yielded r=12.5 mm, g_1=5.7 mm, g_2=2.0 mm, and $\ell_1 = 20$ mm. Figure 1(b) illustrates that although the measured response deviates from ADS simulations for intermediate frequencies, both measurement and simulation are below the target maximum return loss of -10 dB.

The pattern of the disc monopole was also simulated in ADS (not plotted), exhibiting a nearly omni-directional pattern with nulls in the direction of the ground plane, as expected. To obtain a uni-directional pattern, absorber could be placed behind the sensor, which may be desirable for many sensing applications. However, for simplicity the omni-directional pattern was not modified in our experiments.



(a) Dimensions (b) Response

Figure 1: Broadband disc monopole with truncated groundplane: (a) design parameters, (b) simulated and measured return loss

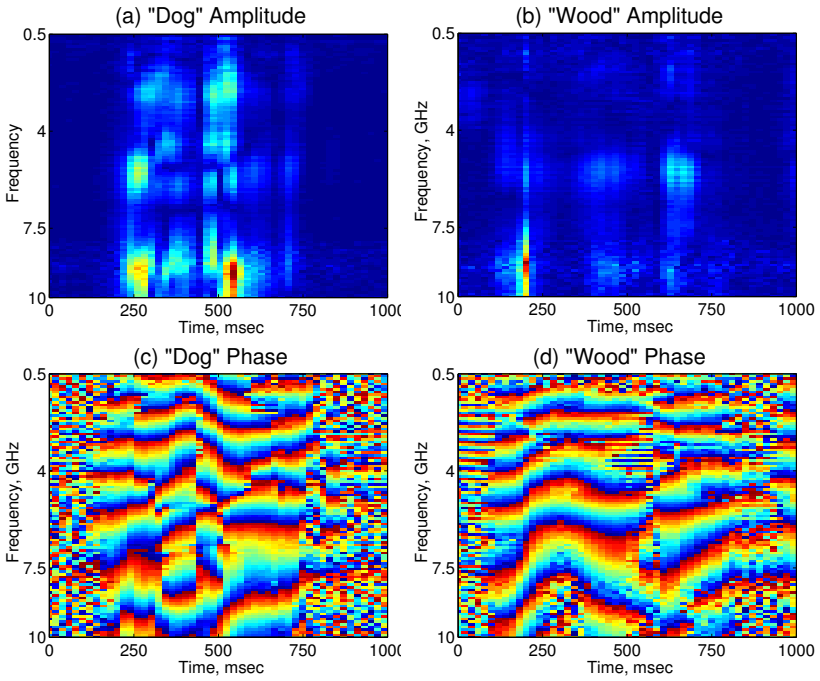


Figure 2: Amplitude and phase of delta response for words *dog* and *wood*

Time-Varying UWB Response Measurement

Tracking of speech production requires transmitted UWB signals to interact with the mouth and throat and the resulting scattered waves to be measured. A simple monostatic radar setup was constructed by mounting the UWB sensor on a prototype headset, positioning the disc monopole approximately 1-2 cm from the mouth of the user.

The monostatic response was measured by connecting the UWB sensor to Port 1 of a Rohde&Schwarz VNB20 vector network analyzer, with a frequency sweep of 101 points over the range of 500 MHz to 10 GHz and a resolution bandwidth of 10 kHz. Time variation of the UWB response was captured using a periodic trigger with a 20 ms period and 100 points. Calibration was performed to remove the effect of the cable up to the UWB sensor. The response was stored on a PC, where S_{kn} refers to the complex reflection coefficient at frequency index k and time point n.

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 $D_{nk} = S_{n+1,k} - S_{nk}$. Figure 2 depicts an example delta response for vocalizations of the words *dog* and *wood*. Note that during the periods of silence at the start and end of the record, there is minimal change, leading 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 reflected in the wideband phase response.

As an initial test of the potential of speech tracking with UWB, a simple speech recognition experiment was conducted, where the UWB response for seven different vocalizations (words) was recorded, representing a minimal dictionary, where the response for word m is $D_{kn}^{(m)}$. Next, the UWB response of a new vocalization is recorded, denoted D_{kn}, and compared with the dictionary. Initially, a sliding

Table 1: Example speech recognition experiment

	Dog	Bat	Test	See	Lab	Bird	Wood	Error
Dog	19	0	0	0	0	0	1	5%
Bat	0	20	0	0	0	0	0	0%
Test	0	0	13	7	0	0	0	35%
See	0	0	1	18	0	1	0	10%
Lab	0	0	0	0	20	0	0	0%
Bird	0	1	0	0	0	19	0	5%
Wood	0	0	0	0	0	0	20	0%

correlator is used, where the estimated word index is

$$\hat{m} = \arg \max_m \max_i \frac{|\sum_{n=1}^{N-i} \sum_{k=1}^{N_F} D_{kn} D_{k,n+i}^{(m)*}|}{\sum_{n=1}^{N} \sum_{k=1}^{N_F} |D_{kn}^{(m)}|^2}, \tag{1}$$

where N and N_F are the number of time and frequency samples, respectively.

Table 1 shows the results of the speech recognition experiment involving the words *bat*, *bird*, *dog*, *lab*, *see*, *test*, and *wood*. After recording the dictionary, each word on the left column is spoken 20 times, and tabulated are the number of times it is matched with each word in the dictionary. The rate of recognition mismatch is quite low (around 5%), except for the word *test*, which is often confused for the word *see*. This is quite encouraging, given the simplicity of the algorithm and that little effort is made to say the words exactly the same each time.

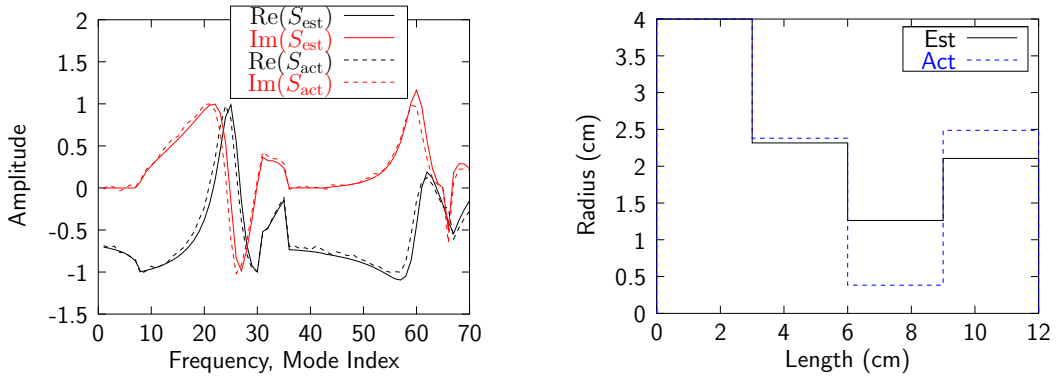
Proposed Waveguide Model

Estimation of the state of vocal tract from the measured return response requires inverse scattering methods for waveguides. The development of direct inverse scattering methods for our problem is challenging, due to the presence of multiple propagating and evanescent modes that contribute to the scattered signal. As a first step, we consider a brute-force reconstruction technique, where the response is compared with a library of simulated responses for all possible parameters at a given level of parameter quantization. The results indicate that although features in the waveguide can be estimated, estimating parameters beyond significant obstructions is quite difficult, possibly requiring transmission (as opposed to reflective) measurements.

The vocal tract is modeled as a cascaded series of perfect electrical conductor (PEC) circular waveguides, where the ith section has common length L and distinct diameter d_i. For the interface from the ith to jth section, the multimode (propagating and evanescent) S-parameters are found using the mode-matching technique [2], where the assumption that the waveguides have a common axis is used, allowing the required integrations to be performed rapidly in closed form.

Estimation of the waveguide parameters from the frequency-dependent multi-mode reflection coefficient is performed by computing a library of responses for all possible combinations of d_i. Given the response, the unknown set of diameters is found by comparing that response against all library entries, and the parameters from the closest match is declared the state of the waveguide.

As a tractable example, a 4-section waveguide is considered, where the diameter of the first section (d_0) is assumed to be known (4 cm) and the cascaded chain is terminated at the right in a short circuit for all modes ($\Gamma = -1$). The possible d_i are quantized at 20 levels ranging from 0 to 4 cm. Figure 3 depicts a typical reconstruction, where the UWB response at 35 frequencies ranging from 3 to 10 GHz is



(a) Response (b) Reconstruction

Figure 3: Example reconstruction using a library lookup based method: (a) matched response, (b) estimated waveguide parameters

Table 2: Percent Error Performance

(a) vs. Upper Frequency (2 modes)

f	6 GHz	10 GHz	20 GHz
Mean($\mathrm{Err_a}$)	33.1	23.3	19.4
Std($\mathrm{Err_a}$)	19.7	18.3	17.0
Mean($\mathrm{Err_w}$)	15.6	8.2	5.1
Std($\mathrm{Err_w}$)	11.2	6.8	3.9

(b) vs. Number of Modes (10 GHz)

Modes	1	2	10
Mean($\mathrm{Err_a}$)	21.7	29.7	25.1
Std($\mathrm{Err_a}$)	18.4	19.1	18.8
Mean($\mathrm{Err_w}$)	7.1	10.4	7.7
Std($\mathrm{Err_w}$)	5.9	7.9	6.3

computed, mode-matching computations are performed using 10 modes, and additive white Gaussian noise (AWGN) is added to obtain a measurement SNR of 30 dB.

The accuracy of this method is investigated by computing the weighted reconstruction as

$$\mathrm{Err} = \sqrt{\frac{\sum_{i=1}^{N_s} |w_i(d_i - \hat{d}_i)|^2}{\sum_{j=1}^{N_s} |w_j d_{\max}|^2}}, \tag{2}$$

where N_s is the number of sections to be estimated (one less than the physical number of sections), and w_i is the weight of the ith section. The *absolute* error $\mathrm{Err_a}$ is computed assuming $w_i = 1$, placing equal error on all sections. The *weighted* error $\mathrm{Err_w}$ is computed assuming $w_j = \prod_{i=1}^{j-1} d_i/d_{\max}$, where $d_{\max} = \max_i d_i$, reducing the weight for sections after obstructions, which are expected to be difficult or even impossible to estimate.

Table 2 lists the mean and standard deviation of the errors for 100 Monte-Carlo simulations, where various frequency ranges and number of compared lowest-order modes are considered. The results indicate that whereas increasing the upper frequency can increase the estimation performance, using more than the lowest order mode does not appear particularly helpful. This latter effect is likely due to the fact that only one mode is not evanescent for the size of the waveguide under consideration.

References

[1] E. M. Staderini, "UWB radars in medicine," *IEEE Aerospace and Electronic Systems Magazine*, vol. 17, pp. 13–18, Jan. 2002.

[2] A. Wexler, "Solution of waveguide discontinuities by modal analysis," *IEEE Trans. Microwave Theory Tech.*, vol. 9, pp. 508–517, Sep. 1967.