ACOUSTIC MEASUREMENTS OF THE VOCAL-TRACT AREA FUNCTION: SENSITIVITY ANALYSIS AND EXPERIMENTS

• *Hani Yehia* ⋆ *Masaaki Honda and* • *Fumitada Itakura*

• School of Engineering, Nagoya University, Nagoya, JAPAN
⋆ NTT Basic Research Laboratories, Atsugi, JAPAN

ABSTRACT

A method to determine the vocal-tract cross-sectional area function from acoustical measurements at the lips is analyzed here. Under the framework described by Sondhi and Gopinath (1971) and implemented by Sondhi and Resnick (1983), a sensitivity analysis of the vocal-tract area function, derived from the impedance or reflectance at the lips is performed. It indicates that, in the ideal case, the area function is not heavily affected by random distortions of the impulse response at the lips. Simulations and real measurements show that the method works relatively well, except for regions behind narrow constrictions. In this case, an excitation pulse with high energy, as well as a fine sampling, proved to be important. The excitation used here is a time stretched pulse. It produces an excitation with high energy without the necessity of a high power sound generator device.

1. INTRODUCTION

The acoustic behavior of the human vocal-tract depends basically on its cross-sectional area function[6], at least during the production of voiced sounds. This reason is enough to make the determination of human vocal-tract area function, during the speech production process, an important problem for speech science.

This problem can be approached by a number of different forms: Reconstruction from 3-D CT data[11, 2] is possible, but limited to static positions, since the scanning time (with presently available technology) is not inferior to several seconds. Estimation from 2-D midsagittal profiles[2] is also possible, but the error due to the 2-D to 3-D mapping is only reasonable. In addition, if the 2-D data are obtained by MRI, then scanning time prohibits data acquisition at a good frame rate. If cineradiography is used, then x-ray dosage problems must be taken into account. When x-ray microbeam, electromagnetic, or ultrasonic techniques[10] are used, again, only partial information about the area function is obtained. In these cases, it is possible to get the remaining necessary information by combining the image data with acoustical information present in the speech signal[4].

The estimation of the area function, based only on the speech signal can not be accomplished[7]. The information present in the speech signal can, however, be combined with prior information (given by positional and dynamical characteristics of the human vocal-tract) to allow the estimation of the vocal-tract geometry in terms of some parameter set, from where the area function can be derived[12, 3, 5]. However, in order to obtain such prior information, (at least) one of the above cited methods, or the method examined here, has to be used.

An alternative acoustic method is to derive the area function from the vocal-tract impulse response at the lips. A procedure to realize it was described by Sondhi and Gopinath[9], and implemented by Sondhi and Resnick[8]. Some advantages inherent to this method are: it allows direct estimation of the area function (in contrast with imaging methods), acquisition of information at a frame rate of several frames/sec. (in contrast with MRI), and absolute safety for the subject (in contrast with x-ray). When compared to transfer function based methods[7], it allows independency of information about glottal source and tract length, low degradation due to the lack of high frequency information, and reasonable correction for losses. The disadvantages are mainly related to the accuracy of a practical implementation of the method.

The objective of this paper is to extend the analysis done in[8] by a study of the sensitivity of the estimated areas due to distortions of the measured impulse responses at the lips. Simulations are realized with the purpose of obtaining a good estimation of the accuracy of the method. Real measurements are also performed.

2. THEORY

In[9] it is shown that if $f(x,t)$ is the solution of the integral equation (written in a normalized unit system where the sound velocity c, the air density ρ, and the lip area A_0 are all unity.)

$$f(x,t) = \frac{1}{2}\int_{-x}^{x} h(|t-\tau|)f(x,\tau)d\tau, \quad |t| \le x, \qquad (1)$$

where x is the distance from the lips and t is time; then the vocal-tract area function is given by

$$A(x) = f^2(x,x). \qquad (2)$$

The kernel $h(|t-\tau|)$ of the integral equation is defined from the impulse response at the lips, which is of the form

$$\hat{h}(t) = \delta(t) + h(t), \qquad (3)$$

and expresses the time domain relationship between pressure and volume velocity at the lips

$$P(0,t) = \int_{-x}^{t} U(0,t)\hat{h}(t-\tau)d\tau. \qquad (4)$$

0-7803-2431-5/95 $4.00

Alternatively, it is possible to solve the problem defined by the following equations[8]

$$\frac{\partial U}{\partial x} + \frac{U(x,x)}{P(x,x)}\frac{\partial P}{\partial t} + \tilde{u} = 0, \quad (5)$$

$$\frac{\partial P}{\partial x} + \frac{P(x,x)}{U(x,x)}\frac{\partial U}{\partial t} + \tilde{p} = 0, \quad (6)$$

$$\tilde{p}(x,t) = \gamma(x)U(x,t), \quad (7)$$

$$m(x)\frac{\partial \tilde{u}}{\partial t} + b(x)\tilde{u} + K(x)\int_0^t \tilde{u}(x,\tau)d\tau = P(x,t), \quad (8)$$

$$\tilde{u}(x,x) = 0, \quad \frac{\partial}{\partial t}\tilde{u}(x,x) = 0, \quad (9)$$

with boundary conditions

$$P(0,t) = 1 + S(t), \quad U(0,t) = 1 - S(t), \quad (10)$$

and the causality condition

$$U(x,t) = P(x,t) = \tilde{u}(x,t) = 0, \quad \text{for } x > t. \quad (11)$$

The area function is then given by

$$A(x) = \frac{U(x,x)}{P(x,x)}. \quad (12)$$

Here, x and t are position along the tract and time in the normalized unit system; $U(x,t)$ and $P(x,t)$ are respectively sound pressure and volume velocity inside the tract; $\tilde{p}$ is the pressure drop per unit length due to viscous losses; $\tilde{u}$ is the volume velocity shunted by the tract walls per unit length; $m(x)$, $b(x)$, and $K(x)$ are respectively the wall mass, resistance, and stiffness along the tract; $\gamma(x)$ is the viscous resistance; and $S(t)$ is the step reflectance at the lips (i.e. reflected pressure when the incident pressure is an unit step).

Hence, if the impulse response or the (step) reflectance can be measured, the area function can, in principle, be obtained. An experiment, as well as simulations based on this method, (and on variations of it) were carried out in[8]. From the results presented there, is was observed that the method works well for the vocal-tract front cavity, but presents problems concerning the estimation of the back cavity (when existent). The reasons for such problems will be examined in the next sections.

3. EXPERIMENTAL APPARATUS

An apparatus, similar to that described in[8] was assembled (see Figure 1). In the system implemented here, the acoustic tube is made of brass, with an an internal diameter of 2.25cm, and a total length of 125cm. A *1 inch* condenser microphone (B & K 4414) is used as speaker; while the used microphone, placed at 65cm from the speaker, is a *1/4 inch* condenser microphone (B & K 4135) with a pre-amplifier. The sampling frequency of the AD converter is 64kHz.

The filter is an anti-aliasing filter which also removes low frequency AC noise and spurious drifts. Its lower and upper cutoff frequencies are respectively 60Hz and 16kHz. Admittedly, the lower cutoff frequency is rather high. It was chosen so due the fact that the microphone presently used as sound source, with a pipe as acoustic load, has a derivative characteristic, which produces a sound pressure that decreases with frequency at 20db/decade. In the future, with a better sound source available, this cutoff frequency can be lowered to 3 or 5Hz. The upper cutoff frequency could have been chosen at any intermediate point between the sound source upper limit frequency (10kHz), and the Nyquist frequency of the AD converter (32kHz).

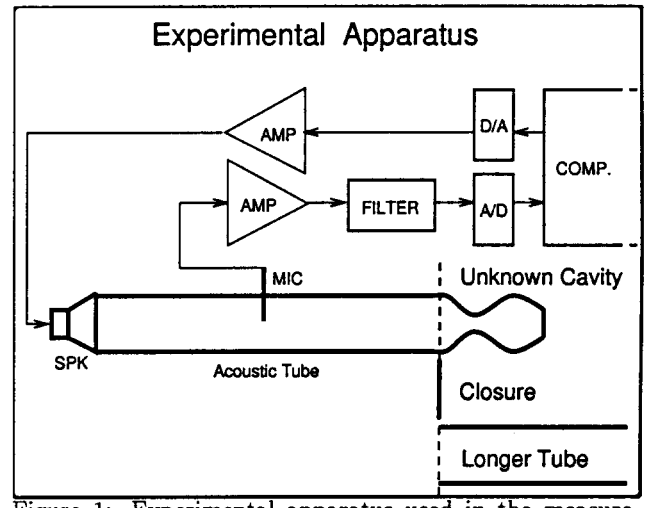

Figure 1: Experimental apparatus used in the measurements.

4. EXCITATION PULSE

In order to estimate the vocal-tract area function, it is necessary to measure the *reflectance* at the lips (from where the *acoustic impedance* and *step reflectance* can be derived[8]). Since the *reflectance* is defined by the reflected signal coming from the vocal-tract when the incident signal is an impulse, the ideal excitation pulse would be an impulse. However, once the sound source has finite power and bandwidth, such an excitation can not be physically realized.

In a practical system, an impulse can be approximated by a very short pulse with power as high as possible[8]. An alternative approach, adopted here, is to use a *time stretched pulse*[1] (TSP). A TSP has basically the same magnitude spectrum of a band-limited impulse. The dif-

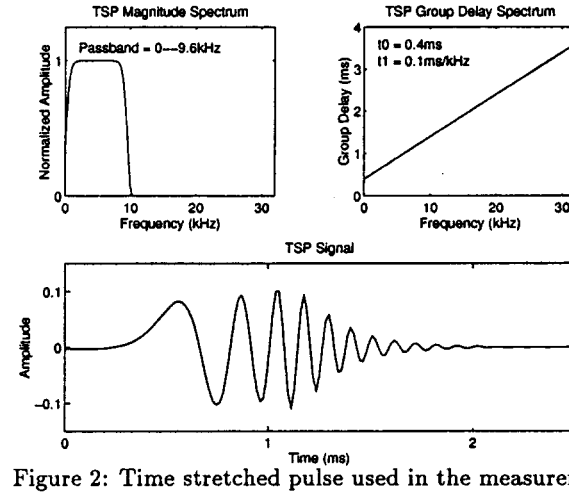

Figure 2: Time stretched pulse used in the measurements.

ference is in the group delay spectrum, which is zero for all frequencies in an ideal impulse, and linear in a TSP. Thus, it is possible to adjust the group delay and distribute the frequency components of the excitation pulse along the time. Therefore, the energy of the pulse is not concentrated in a very short interval any more, and the power requirements of the sound source can be alleviated.

From another point-of-view, in the implemented experiment, the excitation pulse can not be too long, otherwise there would be interference between incident and reflected pulses. Considering the dimensions of the acoustic tube used in the system, an appropriate excitation pulse should be contained in a time interval of about 2ms.

An illustration of the pulse used in the measurements is shown in Figure 2. The bandwidth is limited to 10kHz because the mathematical model used to derive the area function uses the hypothesis of plane wave propagation, which can not be ensured for higher frequencies. (In fact, the theoretical limit for a tube with a diameter of 2.25cm is about 7.6kHz.) If a wider bandwidth is used, then either the pulse becomes prohibitively long, or less energy is given to the frequency range of main interest.

5. THE MEASUREMENT PROCESS

The process used to estimate the area function is illustrated in Figure 3 and described in detail in[8]. The first step is to determine the reflectance. Looking at Figure 1, it is possible to see that, for a given incident pulse, the reflected pulse of an unknown cavity (c), acquired at the microphone, is the result of the convolution of the reflectance of the cavity with the reflected pulse obtained when the acoustic tube is ended by a hard termination (a). The reflectance can then be obtained by a deconvolution procedure. In order to cancel deterministic undesirable signals, the response obtained with a longer tube (b) is subtracted from both hard termination (a) and cavity (c) reflected signals. The results are shown in figures (d) and (e), respectively. The reflectance, obtained from the deconvolution of (d) out of (e), is shown in Figure (f).

Next, the first millisecond of the step reflectance (h) is obtained by direct integration of the first millisecond of the reflectance (g). The first 17cm of the cavity can then be estimated by a numerical solution of the system of equations described in section 2. The square root of the area obtained for this example is shown by the solid line in (i), while the real cavity is represented by the dashed line. A brief analysis of the result indicates a good match up to the second subcavity, and large discrepances in the last (back) subcavity. There are two groups of causes for such discrepances. One is related to the formulation of the model that describes the acoustics of the system[8]. It includes losses, plane wave assumption, etc. The other is related to the accuracy of the measurements, and affects mainly the back cavity, which is reached by only a small fraction of the energy of the incident pulse. Moreover, the estimation error along the cavity is cumulative, since the estimation of the area at a given point depends on the area estimated before it. Some aspects of this fact will be seen in the next section.

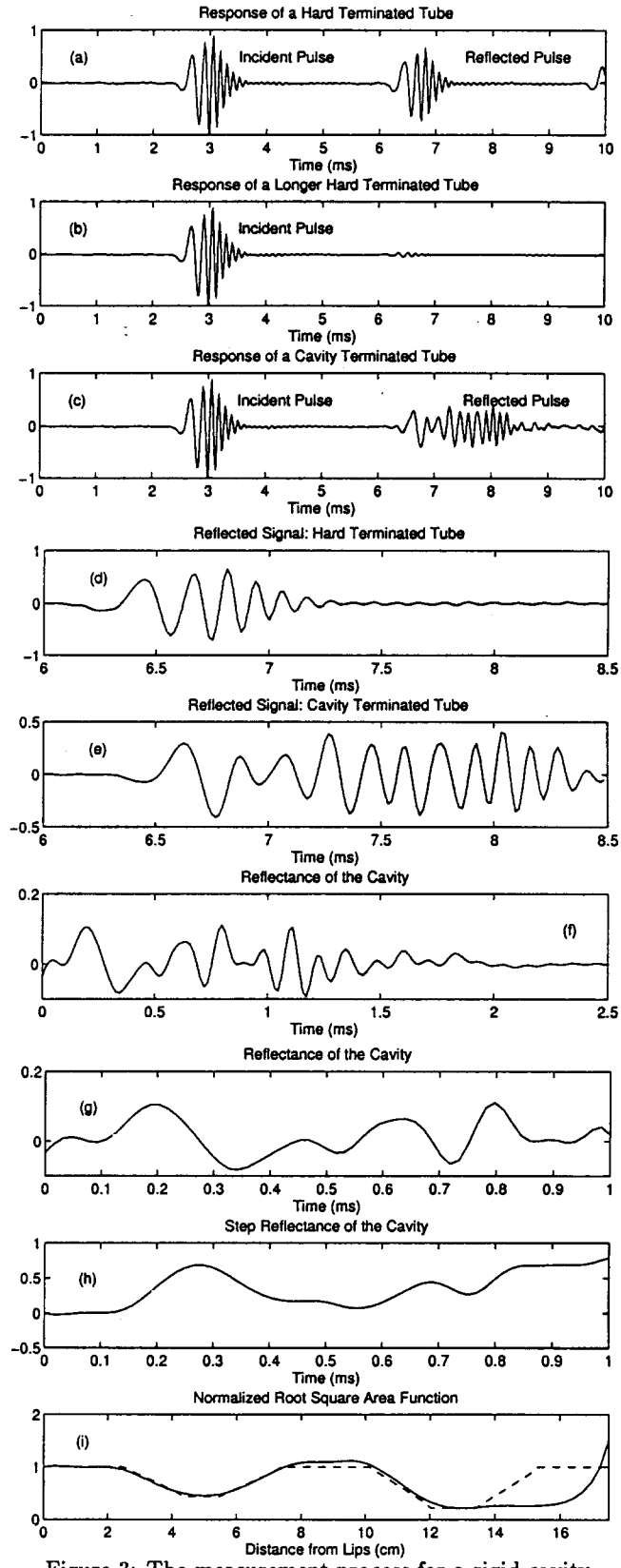

Figure 3: The measurement process for a rigid cavity.

6. SENSITIVITY ANALYSIS

In order to understand some factors that have influence on the accuracy of the method, basic simulations are carried out here. Figure 4 shows sampling effects. In this figure, the solid lines show an area function and an impulse response that are analytically related[9]. The dotted and dashed lines show the areas obtained from simulations where the impulse response was sampled at 50kHz and 200kHz, respectively. It can be seen that a high sampling frequency is particularly important for the back cavity estimation.

Figure 5 illustrates the problem of corruption by noise of the impulse response. It is interesting to note that, if the sampling rate is sufficiently high, the area function is quite insensitive to a zero mean additive random noise corruption of $h(t)$. (It happens due the fact that $A(x)$ is directly related to an integral of $h(t)$.) Unfortunately, during the measurements, noise is many times result of spurious reflections of the signal and, hence, is not really random noise.

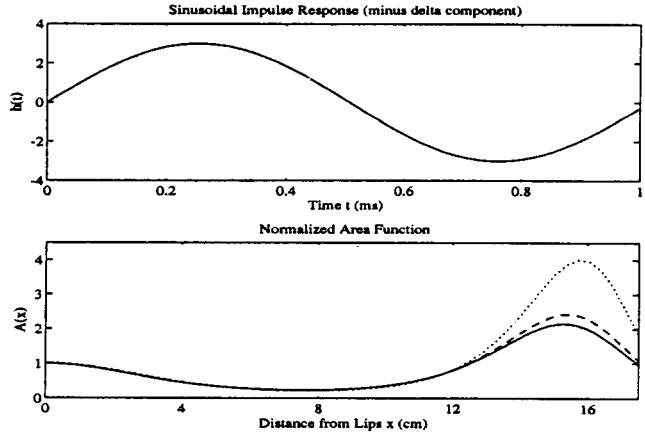

Figure 4: Theoretical (solid line), and simulations at sampling rates of 50kHz (dotted line) and 200kHz (dashed line), of the area function correspondent to an impulse response of the form $\hat{h}(t) = \delta(t) + h(t)$ where $h(t) = 3\sin(\pi t)$.

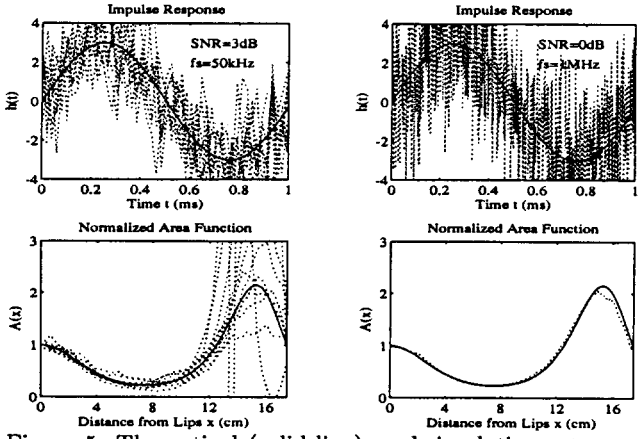

Figure 5: Theoretical (solid line), and simulations at sampling rates of 50kHz (left) and 1MHz (right), of the area function correspondent to the same impulse response of Fig.1. Here, $h(t)$ is corrupted by gaussian white noise. The SNR is 3dB in the left (10 realizations are shown), and 0dB in the right (one realization is shown).

7. CONCLUSION

A method for direct acoustical measurement of the human vocal-tract area function is studied here. The method is based on the relationship between the reflectance (or impulse response) at the lips and the area function. In the experimental procedure, the use of a *time stretched pulse* has shown to be effective to obtain the reflectance. The example given in section 5 shows that the method works relatively well, but also shows the difficulty to obtain accurate measurements of "back cavities."

In the sensitivity analysis, the importance (especially for the back cavity) of a fine sampling of the measured impulse response is shown. Another point observed here is the low sensitivity of the area to zero mean noise corruption of the impulse response at the lips (if the sampling rate is high enough).

At this moment, preliminary experiments are being carried out with human subjects and compared with the results of lossy cavities. The main points that still need to be improved are the sound source, the coupling between mouth and the acoustic tube, and modelling of losses.

8. REFERENCES

[1] N. Aoshima Computer-generated pulse signal applied for sound measurement. *JASA*, Vol. 69, No. 5, pp. 1484–1488, 1981.

[2] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye. Analysis of vocal tract shape and dimensions using magnetic resonance imaging. *JASA*, Vol. 90, No. 2, pp. 799–828, 1991.

[3] G. Bailly, R. Laboissière, and J. L. Schwartz. Formant trajectories as audible gestures: an alternative for speech synthesis. *J. Phon.*, Vol. 19, pp. 9–23, 1991.

[4] M. Honda and T. Kaburagi. Estimation of articulatory-to-acoustic mapping using artificial neural network model (in Japanese). SP-92 144, IEICE, 1993.

[5] J. Schroeter and M. Sondhi. Speech coding based on physiological models of speech production. In *Advances in Speech Proc.*, pp. 231–268. Marcel Decker, 1991.

[6] M. M. Sondhi. Model for wave propagation in a lossy vocal-tract. *JASA*, Vol. 55, No. 5, pp. 1070–1075, 1974.

[7] M. M. Sondhi. Estimation of vocal-tract areas: The need for acoustical measurements. *IEEE ASSP*, Vol. 27, No. 3, pp. 268–273, 1979.

[8] M. M. Sondhi. The inverse problem for the vocal-tract: Numerical methods, acoustical experiments, and speech synthesis. *JASA*, Vol. 73, No. 3, pp. 985–1002, 1983.

[9] M. M. Sondhi and B. Gopinath. Determination of vocal-tract shape from impulse response at the lips. *JASA*, Vol. 49, No. 6, pp. 1867–1873, 1971.

[10] M. Stone. A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data. *JASA*, Vol. 87, No. 5, pp. 2207–2217, 1990.

[11] C. S. Yang and H. Kasuya. Accurate measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects. In *Proc. ICSLP*, 1994.

[12] H. Yehia and F. Itakura. Determination of human vocal-tract dynamic geometry from formant trajectories using spatial and temporal Fourier analysis. In *Proc. IEEE ICASSP*, 1994.