CAN THE AREA FUNCTION OF THE HUMAN VOCAL TRACT BE DETERMINED FROM THE SPEECH WAVE?

Hans Werner Strube

Drittes Physikalisches Institut, Universität Göttingen, Germany

Introduction

There has been considerable interest in determining the area function of the vocal tract during the speaking process. For vowel-like sounds, the area function is the link between the position of the articulators and the acoustical quantities of the speech wave. Even if the articulatory parameters have been measured by other means, their relation to the area function is not clear a priori; thus an independent area determination, best from the speech signal itself, could help to clarify this relation. If, on the other hand, reasonable assumptions concerning this relation are available, the articulatory parameters can be determined much more easily from the speech wave (via the area function) than by direct measurements.

In the last several years, attempts to determine the area function have become centered around Wakita's method. With appropriate preemphasis they yield reasonable-looking results. Nevertheless, I feel that the theoretical difficulties with such methods should be reiterated (e.g., what preemphasis is "appropriate"?) hence the question mark in the title. In the first part of this paper, we shall recapitulate the theoretical background and obtain some new results to illuminate the scope of the difficulties. In the second part, experiments with parametric models will be described.

I. DIFFICULTIES OF PRESENT METHODS

Theoretical Background

The determination of the area function is a special problem of transmission-line synthesis. We need a physical model of the line (tube, tract) and appropriate acoustic measurements. Most of the original methods assume the tube as hardwalled, without internal losses, with a hard glottis (current source) - all of which are unrealistic assumptions. The tube may be continuous or segmented. The frequency-domain methods employ, for example, two sets of resonance frequencies for different boundary conditions, the pole and zero frequencies of the input impedance at the glottis or lip end, the poles and their residues, or the pole (formant) frequencies and bandwidths of the transfer function. Since only a small number of these quantities can be measured in the range up to about 4 kHz (above which the models do no longer apply), artificial high-frequency information has to be filled in for some methods; others immediately yield a smoothed area function when working with a finite number. In any case, resolution is no better than 2 cm for frequencies up to 4 kHz.

In frequency-domain methods generally, the tube length and the load impedance have to be known. However, the area function may also be derived in the time-domain without this knowledge. For
the continuous-time (and -tube) case, Sondhi and Gopinath\textsuperscript{5})
proved that a piece of the area function $S(x)$ can be obtained
from a piece of the pressure response to a volume-velocity im-
pulse at the tube input. A slightly more general discrete-time
analogue was proved by Atal\textsuperscript{6}) for a segmented tube of segment
length $c/2f_g$ ($f_g$ = sampling frequency):

**Theorem 1:** Let the input pressure $p_n$ and volume-velocity $u_n$ be
zero for $n < 0$. If two of the three $N$-tuplets $u_{n-1}$, $p_{n-1}$, $g_n$
($n = 1$, ..., $N$) are known, the third can be determined.

Unlike the complicated continuous case, here very simple recur-
sion formulas result from the transmission difference-equations
(equivalent to the frequency-domain method of Paige and Zue\textsuperscript{7})).
Obviously there is no dependence on tube length and termination.

Since it is desired to work with the speech signal itself, one
has to start from a transfer function of the tract rather than
from an input impedance. However, knowing the load impedances,
the input impedances may be reconstructed from the transfer
function. The lip impedance is obtained by a decomposition of
the denominator-polynomial of the volume-velocity transfer
function\textsuperscript{8),9}), but good knowledge of tube length and radiation
impedance is required. More advantageous is the use of the
glottal-end input impedance, whose time-domain equivalent, ac-
 accordance to Atal\textsuperscript{6}), can be determined from the transfer prop-
erties:

**Theorem 2:** Let a lossless tube be terminated by a load admitt-
ance $Y(\omega)$ with constant real part $G$ (or in the discrete-time
case, $Y(z) + Y(z^{-1}) = 2G$). Let $Z_1(t)$ be the input-pressure re-
 response and $R(t)$ the ACF of the output-pressure response to a
delta- (or unit-) impulse of input-volume-velocity. Then

$$Z_1(t) = \begin{cases} 0, & t < 0, \\ GR(0), & t = 0, \\ 2GR(t), & t > 0. \end{cases}$$

(Originally, in his San-Diego-talk\textsuperscript{6}), Atal considered only a
real $Y = G$.) Combining this with Theorem 1 shows that a piece
of the area function can be determined from a piece of the ACF
$R(t)$, since the radiation admittance, approximately an RL par-
allel circuit, is of the form required in Theorem 2. As $G$ is
proportional to the lip area, a common factor remains unknown
(the same holds for most other methods).

**Importance of proper model: Atal vs. Wakita**

Atal\textsuperscript{6}) also showed that (for discrete time) the area function
can be obtained from the ACF $R(t)$ without intermediate calcu-
lation of $Z_1(t)$. Using the Levinson algorithm, predictor polyno-
mials $A_n(z)$ are recursively computed from the ACF, and the
highest coefficients $a_n$ (i.e., $-k_n$, PARCOR-coefficient) are inter-
preted as reflection coefficients $r_n$ between the tube segments,
in the order from glottis to lips. This looks much like Waki-
ta's method\textsuperscript{1}), and in fact, both methods differ only in fre-
quency characteristics, since Wakita uses the volume-velocity
transfer function, Atal the volume-velocity-to-pressure trans-
fer function. Output pressure is approximately the time derivative of volume-velocity because of the highly inductive load. If in Atal's model the radiation admittance were real and constant, the methods would become identical. Nevertheless, the results have to be interpreted in a completely different way, since the boundary conditions of Wakita's tube model are different from Atal's (and all previous authors): the radiation impedance is zero and the losses lumped into a constant resistive load at the glottal end. So these two nearly identical methods offer a possibility to demonstrate the effects of overall frequency characteristics and boundary conditions on the area function obtained. For the differences between Atal and Wakita the effects are very crude: the reflection coefficients are obtained in opposite order. Thus, when the logarithm of the area function is plotted, both methods differ by a rotation (of the log-area curve) by 180°. This crude effect is mainly due to the choice of the underlying physical tube model; for (because of the unrealistic model assumptions) the difference in the frequency characteristics turns out to be much smaller than originally expected. So one is left mainly with a difference in interpretation of the results: has the log-area curve to be rotated by 180° or not? This is often difficult to decide when both possibilities look reasonable (a symmetry that was already exploited by Schroeder and Mermelstein).

In fact, both models are highly idealized. In a real vocal tract, the losses are composed of radiation, viscosity and heat-conduction losses, all increasing with frequency, and wall-vibration and glottal losses, both effective at low frequencies. Of course, most of these losses are not lumped at one end of the tube. For a wide mouth opening as in /a/ the radiation loss dominates, so one might expect Atal's model to be superior, whereas for narrow constrictions at the lip end as in /i/ and /u/, internal and wall losses are important and thus Wakita's model might be more reasonable. Experiments actually point in this direction. E.g., the lip constriction in /u/ can only be obtained by Wakita's method; /a/ comes out better with Atal's method. In the tests, synthetic signals have been employed derived from Fant's area functions terminated by an RL-parallel circuit at the lips and by an RL-series shunt at the glottis \((R = 25 \text{ g cm}^{-4} \text{s}^{-1}, L = 0.02 \text{ g cm}^{-4})\), the latter crudely representing the wall impedance. Internal air losses have not been modelled. Even when in Atal's method the exact ACF is known, the finite wall impedance (increasing the 1st formant frequency and bandwidth) falsifies the result considerably, especially for narrow mouth opening - as expected. In Fig. 1 such area functions (segment length 1 cm, \(f_s = 17.65 \text{ kHz}\)) are shown: they are widened and shifted towards the lip end, except when the mouth is wide open (1st formant high). This can partly be compensated by weaker preemphasis.

**Importance of frequency characteristics**

The importance of appropriate preemphasis is well known and has been emphasized by Wakita. Further improvements were attempted by Nakajima et al. However, all such approaches are highly empirical - see e.g. the investigations by Tanaka. The preceding section already made clear how difficult a theoretical foundation of proper preemphasis will be. Indeed, even the
shortcomings of the underlying model (having different effects for different vowels) must partly be compensated by preemphasis; thus preemphasis actually depends on the area function that is to be determined! Moreover, the glottal frequency characteristic varies even for one speaker. The effects of wrong preemphasis are just as crude as those of a wrong model. Excessive preemphasis has a similar effect as the wall impedance: shifting and widening towards the lip end (in Atal's interpretation). For two simple cases, we can give quantitative results:

1. Uniform tube, resistive load. With +6 dB/octave wrong preemphasis, in Wakita's method $S_n \propto 1/n(n+1)$ is obtained instead of $S_n = \text{const.}$ (Atal's method: $S_n \propto n(n+1)$.)

2. Any tube, either model; all-pole transfer function. With -6 dB/octave wrong preemphasis, the area function approximately becomes its reciprocal and is shifted by one segment.

Experiments with synthetic and natural vowels seem to indicate that the optimal preemphasis (applied to the free-field sound pressure) ranges between +6 and +12 dB/octave.

**Difficulties in measuring the transfer function**

So far, we have assumed that the transfer function or corresponding ACF of the tube is exactly known. Actually, it has to be estimated from the speech wave. This means, that the unknown glottal input spectrum must be removed. First, the input spectrum affects the frequency characteristics; second, the spectral envelope is defined only on discrete points (multiples of the fundamental frequency) so that small bandwidths are difficult to estimate. When an ACF or power spectrum is calculated using a finite time window, the spectrum is smeared. The same holds for the PARCOR method or when cepstral smoothing is applied to obtain the spectral envelope. Another smearing effect results from the periodic variation of glottal impedance; e.g., the first formant frequency may differ by 200 Hz between the
closed-glottis and open-glottis intervals; also the bandwidths vary\(^3\). All these effects yield much too large measured bandwidths as compared to theory or direct measurements with laryngectomized subjects. However, the bandwidths are essential for the determination of the area function; without their knowledge, an infinite set of solutions is possible. As a result, some authors have even given up the use of measured bandwidths and instead employ assumed values\(^3\),\(^4\) or other criteria to obtain uniqueness\(^3\),\(^4\). Others use empirical resonance enhancement\(^4\).

A more fruitful approach to bandwidth measurement may be Schroeder's and Atal's "covariance method" of linear prediction, allowing very short analysis intervals. If closed-glottis intervals exist, analysis on such an interval should allow to derive the exact hard-glottis transfer function or ACF. For an idealized signal, the method is even hardly sensitive to differentiation. In practice, however, it is highly sensitive to wrong choice of predictor order, noise and a nonvanishing input signal of the tube. Even the positivity of bandwidths is not insured. The closed-glottis interval is difficult to determine and easily "smeared out" by the recording devices.

Conclusions

The direct computability of the area function from the speech signal is impeded by many difficulties. Deviations of the underlying physical model from reality and a wrong overall frequency characteristic may have catastrophic effects on the results, and even the measurement of the relevant acoustic quantities is loaded with much uncertainty. To put an end to so much pessimism, we stress that reasonable-looking results are indeed obtainable by appropriate methods; even practical applications, e.g., for teaching the deaf\(^5\), are being pursued. Our aim was to recall the fact that, at present, theory still has to be supplemented by plenty of empirical trial-and-error. One usually cannot tell whether or not a single result is the "true" area function.

II. EXPERIMENTS WITH PARAMETRIC MODELS

General

The only path to theoretically more reliable results seems to be the use of more realistic vocal-tract models. However, this will in general not lead to direct computational methods. For example, (very complicated) direct methods have been derived\(^6\) only for two special cases of distributed loss or wall impedance. So one is left with iterative methods: the tract model is expressed by several parameters (purely mathematical or articulatory) which are determined by an iterative fitting of acoustical model quantities to the corresponding measured values. Two such models will be employed here; the first is a simple mathematical one, the second a complicated articulatory one. In either case, the tube itself has segments of 0.5 cm (i.e., \(f_S = 35.3\) kHz); the segment length or sampling frequency is here no more connected to the limiting signal frequency (3 or 4.4 kHz). Radiation impedance is modelled as a lip-area dependent RL-parallel circuit, the wall impedance as a constant.
RL-series shunt (see p. 3) at the glottal end, using the bi-
linear z-transform. Internal air losses are not included since
their $\sqrt{f}$-dependence is difficult to express in a simple dis-
crete-time formulation.

Synthetic vowels for experiments have been generated from
Fant's area functions with glottal pulses of the form

$$u(t) = \begin{cases} \sin^2(\pi t/204), & t = 1, \ldots, 102, \\
\sin(\pi(143-t)/82), & t = 102, \ldots, 143, \\
0, & t = 143, \ldots, 256 \text{ (period)}. 
\end{cases}$$

Natural speech sounds have been recorded by a pressure micro-
phone (1/2 inch, condenser type) in an anechoic room, differen-
tiated, low-pass filtered at 7 kHz (48 dB/Oct.) and directly
A/D-converted. In both cases, further limitation to 3 or 4.4
kHz was done digitally by a linear-phase FIR filter, so "smear-
ing out" of the glottal-closure interval was minimal.

Fitting procedure

After several, not very successful, attempts to fit logarithmic
power spectra, it was decided to make use of the above-men-
tioned advantages of prediction on a closed-glottis interval
(for its determination, see Strube\textsuperscript{9}, \textsuperscript{17}). The uncertainties of
the covariance method can be circumvented here by constraining
the predictor to be compatible with the model transfer function.
Two approaches were considered.

1. "Denominator method". The denominator of the transfer func-
tion itself is taken as a predictive inverse-filter and the
prediction error post-filtered by an integrator (works better
than post-filtering by 1/numerator of transfer function), then
the mean is set to zero. The sum of squares of the remaining
error signal is minimized by adjusting the model parameters.
Since no frequencies above $f_s/8$ are present, only every fourth
error-signal value has to be considered.

2. "ACF method". The ACF corresponding to the model transfer
function is low-pass filtered at 4.4 kHz, down-sampled at $f_s/4$,
and a predictor calculated by the Levinson algorithm. The sum
of squares of the prediction error of the (down-sampled) signal
is minimized by adjusting the model parameters.

In either method, 20 samples of the error signal on a glottal-
closure interval were taken. The iteration program for these
fitting procedures was a general-purpose program using Mar-
guardt’s method\textsuperscript{18}, which combines the safety of the gradient
method with the fast final convergence of the Newton technique.

Cosine-series model

The first model represents the log-area function as a cosine-
series expansion, as in the Schroeder-Mermelstein papers\textsuperscript{3}, \textsuperscript{4},
but with a four-segment larynx-tube attached to the glottal end:
Fig. 2: Fitting of Cosine-series model to synthetic (left; originals dashed) and natural vowels /i, e, a, o, u/.

Fig. 3: Vowel transition /ɔɪd/, frame rate 30 s⁻¹. Result of one frame = starting value for next frame.

\[
S_n = \exp \sum_{j=0}^{k-1} b_j \cos \frac{j(n-5)}{N-5}, \quad n = 5, \ldots, N, \\
S_n = 3 + (S_5/4 - 3)(n - 1/2)/4, \quad n = 1, \ldots, 4.
\]

Unfortunately, the tube length \( N \) has to be known. In practice the length values from Pant's area functions were taken, length variation has no severe effects. The "denominator method" worked considerably better than the "ACF method". Results for synthetic (left, \( k = 7 \)) and natural vowels (right, \( k = 6 \)) /i, e, a, o, u/ with 3 kHz limiting frequency are shown in Fig. 2. When iteration starts from a uniform tube (\( b_j = 0, j > 0 \)), the \( b_j \)
with $j$ even have only second-order effects at the beginning; thus convergence is greatly improved by fixing these coefficients at zero (i.e., $S_{n-5} = S_{n-15-n}$) during the first four or five iteration steps. The results look good except that, for natural vowels, $b_0$ is systematically too large. However, this quantity is actually not very important (merely scaling the area values) and ill-defined, mainly dependent on the proper choice of wall impedance (which was assumed known). Further, for natural /u/ the lip constriction is not obtained. This may be due to the lack of air losses in the model, which become important for narrow constrictions and low radiation damping.

The method was also applied to a vowel transition /ora/ at a frame-rate of 30 per second (Fig. 3). The original supposition that using the result of one frame as the starting value for the next one would be advantageous was disproved. Unfortunately, the results depend somewhat on the starting values; even complete failures may occur.

**Articulatory model**

Use of an articulatory model offers the advantage that the set of possible area functions is constrained to be compatible with anatomy, thus reducing ambiguities of solutions. On the other hand, not much is learnt about the relation between articulatory parameters and area function beyond what is already contained in the model. Whereas other authors tried to fit an articulatory model to an area function obtained by "direct" methods [11], here the model is fitted immediately to acoustic data and the area function is merely a byproduct. The model employed is a reconstruction by the author of a preliminary model by Mermelstein, Maeda and Fujimura (1970, unpublished?). Its parameters are jaw angle, lip rounding (both determining lip shape), and height and position of tongue hump, where the tongue shape is described in a rotating coordinate system fixed to the jaw. Coordinates are cartesian in the pharynx and mouth regions and polar in the velar region. When the polar part is rectified, the tongue contour consists of sinusoidal and straight-line pieces. The total length of the resulting area-function is a continuous function of the model parameters, consequently an interpolation for nonintegral segment number had to be found [9,19].

The method was applied only to synthetic vowels. As Fant's area functions cannot be reproduced accurately by the model, vowels have also been synthesized using the model itself. In any case, the jaw angle has to be known and fixed during the iteration. This was expected since there is much ambiguity between jaw and tongue parameters, corresponding to the well-known possibility of articulatory compensation. Both the denominator and the ACF method often give very good results especially for the model-derived vowels, but sometimes the procedures fail - unsystematically depending on circumstances as vowel, upper limiting frequency, parameter starting values etc. Effects of finite numerical accuracy seem to play some part due to the highly complicated transformations in the model routines. The iteration procedure using this model could not be made sufficiently reliable for practical application to real speech. Some results for the ACF method are presented in Fig. 4. In this special ex-
Fig. 4: Fitting of articulatory model to synthetic (model-derived) vowels. Top: log-area functions (dashed: originals), bottom: mid-sagittal section, lip length, lip shape.

ample, the method failed for /i/.

Conclusion

A new way of fitting arbitrary parametric (mathematical or articulatory) models of the vocal tract to acoustic speech data is presented. Roughly speaking, the method is based on linear prediction under model constraints. By working with the closed-glottis intervals, the uncertainties associated with the unknown source spectrum and glottal impedance are greatly reduced. The models can be chosen as realistic as desired since no simple "direct" algorithms are to be derived from them. This is paid for by the necessity of having to use slow iteration procedures which may fail to converge to the correct results in some cases.

References