



Vocal Tract and Area Function Estimation with both Lip and Glottal Losses

Kaustubh Kalgaonkar, Mark Clements

Center for Signal and Image Processing, School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, GA, USA 30332-0250

{kaustubh,clements}@ece.gatech.edu

Abstract

Traditional algorithms simplify the lattice recursion for evaluation of the PARCOR's by localizing the loss in vocal tract at one of its ends, the lips or the glottis. In this paper we present a framework for mapping to pseudo areas the VT transfer function with no rigid constraints on the losses in system, thereby allowing losses to be present at both the lips and glottis. This method allows us to calculate the reflection coefficients at both the glottis (r_G) and the lips (r_{Lip}).

The area functions obtained from these new PARCOR's, have better temporal (inter-frame) and spatial (intra-frame) predictability.

Index Terms: Vocal tract area function, PARCOR's, glottal reflection coefficient.

1. Introduction

Characteristics of the vocal tract (VT) have long been a topic of interest and fascination with engineers and scientists. Previous work [1, 2, 3] used X-Ray or MRI techniques for direct estimation of VT cross-section areas. These methods produce accurate estimates of VT areas, but are not very practical.

Indirect estimation of VT areas from acoustic and speech data is also possible. Gopinath and Sondhi in [4] presented a method for estimation of the VT area-function using the acoustic measurements at one of the ends of the vocal tract. Deng et al. [5] presented an algorithm based on measurement from a p-mic to estimate the glottal reflection coefficient and VT areas. Atal [6] and Wakita [7] presented an inverse filter-based method for estimation of VT area-functions using only speech data.

The problem of estimating the areas of the VT can be split into two parts: estimation of the vocal tract transfer function (VTTF) and separating the source characteristics from the VTTF. Estimation of VT areas under specific conditions can be related to the estimation of the PARCOR's [7]. These conditions stipulate that losses in the VT area model be located at one of its ends (glottis or lips).

In this paper we present a framework to estimate the VT area-function, without making any strict assumption as to the location of the loss in the VT. The goals of our model are listed below:

- To better estimate the area-function of the vocal tract based on speech data only.
- To constrain the area-function, to one that can be naturally attained by the human vocal tract.
- To match the LPC filter exactly.

Although the method improves on the traditional approaches, we do not claim to achieve the true VT area-function.

Section 2 provides a brief overview of the uniform lossless tube model and its relation to inverse filtering and lattice formulation. Section 3 presents the framework for estimation of the glottal reflection coefficients and the algorithm for estimation of smoother VT area-functions, followed by Results and Conclusions in Sections 4 and 5 respectively.

2. The Vocal Tract

2.1. Uniform Lossless Tube Model

One widely utilized model of speech production maps the vocal tract as a series of connected uniform lossless tubes with varying areas (S_m). Given a sufficient number of tubes, it is possible to produce resonances matching those produced by a human vocal tract. The VT is modeled with M uniform tubes, each of length x . Detailed discussion on the acoustic tube models and can be found in [6] and [7]. The VT transfer function is defined as

$$H_{VT}(z) = \frac{U_{Lip}(z)}{U_G(z)} \quad (1)$$

where $U_{Lip}(z)$, $U_G(z)$ are the z-transforms of the discretized lip radiation volume velocity ($u_{Lip}(t)$) and glottal source volume velocity ($u_G(t)$) respectively.

Using the standard pressure and volume velocity wave equations, transfer function of the vocal tract can be written as

$$H_{VT}(z) = \frac{0.5(1+r_G)(1+r_{Lip})z^{-\frac{M}{2}} \prod_{m=1}^{M-1} (1+r_m)}{\begin{bmatrix} 1 & r_G \end{bmatrix} \left\{ \prod_{m=1}^{M-1} \begin{bmatrix} 1 & r_m \\ r_m z^{-1} & z^{-1} \end{bmatrix} \right\} \begin{bmatrix} 1 \\ r_{Lip} z^{-1} \end{bmatrix}} \quad (2)$$

where r_m is the reflection coefficient at the boundary of tubes m and $(m+1)$

$$r_m = \frac{S_{m+1} - S_m}{S_{m+1} + S_m} \quad (3)$$

r_G , r_{Lip} are the reflection coefficient at glottis and the lips respectively and S_m is the cross-section area of tube m . Since the filter is passive, $|r_m| \leq 1$ for all m . To calculate the lip and glottal reflection coefficients, it is necessary to know the areas of the lossless tubes along with the lip and glottal impedances. Calculating the impedances from only speech data is difficult. Equation (2) can represent a gain element G , a delay of $-M/2$ samples and an all-pole filter $\frac{1}{D(z)}$

$$H_{VT}(z) = G \left(\frac{1}{D(z)} \right) z^{-\frac{M}{2}} \quad (4)$$

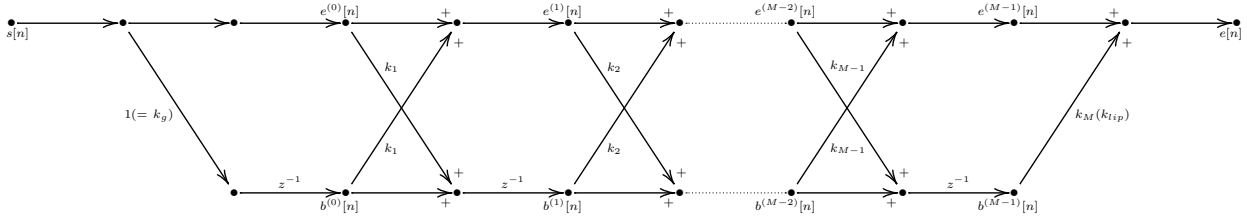


Figure 1: Lattice representation of the VT

2.2. Inverse Filter and Vocal Tract Filter Estimation

Speech production can also be modeled as an autoregressive process with the transfer function given by

$$H(z) = \frac{S(z)}{U_G(z)} = \frac{\tilde{G}}{A(z)} = \frac{\tilde{G}}{1 - \sum_{k=1}^M a_k z^{-k}} \quad (5)$$

Where \tilde{G} is the gain and U_G is the driving glottal volume velocity. The coefficients of the all-pole VT filter can be estimated using *linear predictive analysis* on the speech data [6].

2.2.1. Lattice Formulation

Itakura-Saito demonstrated that the linear prediction coefficients (a_k) can also be obtained using a lattice shown in Figure 1. The PARCOR's k_i are calculated using

$$k_{m+1} = \frac{-\sum_{n=-\infty}^{\infty} e^{(m)}[n]b^{(m)}[n]}{\sqrt{\sum_{n=-\infty}^{\infty} (e^{(m)}[n])^2 \sum_{n=-\infty}^{\infty} (b^{(m)}[n])^2}} \quad (6)$$

Where $e^{(m)}[n]$ and $b^{(m)}[n]$ are the m^{th} order forward and backward prediction errors. The recursion is started with $e^{(0)}[n] = s[n]$ and $b^{(0)}[n] = s[n-1]$ ($k_G = 1$), where $s[n]$ is the frame of speech.

The predictor polynomial $A(z)$ is the impulse response of M^{th} order lattice shown in Figure 1

$$A^{(M)}(z) = [1 \quad k_g] \left\{ \prod_{m=1}^{M-1} \begin{bmatrix} 1 & k_m \\ k_m z^{-1} & z^{-1} \end{bmatrix} \right\} \begin{bmatrix} 1 \\ k_M z^{-1} \end{bmatrix} \quad (7)$$

2.3. Relation between Lossless Tube and Lattice Models

Equations (4) and (5) suggest strong congruence between the vocal tract models obtained from the inverse filtering (lattice method) and the lossless tube model. Comparing the equations for $A(z)$ and $D(z)$, the two models are equivalent if and only if the glottal reflection coefficient $r_G = 1$ which is true only when the glottis is closed. In that case the PARCOR's are same at the reflection coefficients ($k_m = r_m$).

The area functions of the lossless tubes can be calculated using the reflection coefficients and Equation (3) recursively. The task of finding glottal closure is non-trivial. To simplify the problem strong assumptions are usually made about the source, and the losses in the system. Pre-emphasis is used to counter the effect of the glottal source.

Atal et al. in [6] and Wakita [7] used this equivalence between the lattice and lossless tube model to calculate the reflection coefficients and obtained the area-function. They make two fundamentally different assumption about the losses in the VT.

Atal assumes all the loss at the lips and so the glottal reflection coefficient $k_g = r_g = 1$, where as, Wakita assumes all the loss in the system is at the glottis and the lip reflection coefficient $r_{Lip} = k_{Lip} = 1$.

3. Area Function

The existence of many-to-one relationship between area function and resonances of the VT was rigorously discussed in [8]. The lattice representation of the VT filter Figure 1 is a good tool to understand this multivalued relationship. In theory it is possible to come up with distinct combinations of $|k| \leq 1$ that generate lattices with the same response $A(z)$. Each of these combinations of k 's will however produce a distinct area-function. It is important to note that all these orientations may not be attained by the VT.

Figure 2 shows three VT's orientations producing the same impulse response. The VT profile (a) is generated by assuming all the loss is located at the glottis, (c) is generated assuming all the loss is at the lips. Profiles (a) and (c) are two extreme solutions and there may exist a number of other configuration depending on the distribution of the total loss at the lip and glottal ends. Figure 2-(b) is one such orientation of the vocal tract, where $k_g \neq 1$.

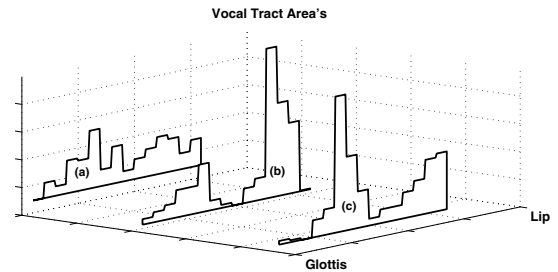


Figure 2: Area functions for a frame of speech with (a) $k_{lip} = 1$ all the loss at glottis. (b) $k_g \neq 1$ & $k_{lip} \neq 1$ loss at both lips and glottis. (c) $k_g = 1$ all the loss at the lips

Itakura-Saito or the Burg's algorithms are some of the methods used to solve for the k 's of the lattice. These algorithms calculate the PARCOR's with the objective to minimize the errors (or some combination of the errors) $e^{(m)}[n]$, $b^{(m)}[n]$ at each stage of the lattice. Due to the simplifying assumptions, resulting area functions lie in the extreme of the solution space (2-(a) and 2-(c)). Also as the solution of the lattice is a recursion, any change in one of the stages creates a ripple that travels down the entire structure. Modifying the value of a reflection coefficient of stage p , (k_p), not only affects the output of that lattice stage $e^{(p+1)}[n]$ and $b^{(p+1)}[n]$ but also changes the reflection coefficients of other stages.

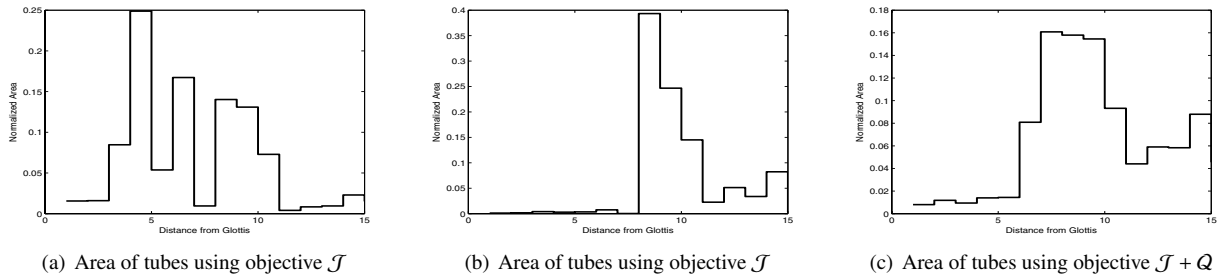


Figure 3: Area function for frames of speech for vowel /a/. (a) and (c) are the areas for the same frame of speech.

3.1. Estimating Reflection coefficient

Traditional methods do not provide the flexibility to distribute the VT losses throughout the system, and cannot be used for estimation of the glottal reflection coefficient. Next we will present a framework that will allow us to estimate all the reflection coefficients (including the lip and glottis) from the speech data only.

We can use the many-to-one mapping between the lattice (PARCOR's)/VT area and its impulse response to our advantage and formulate the estimation of the reflection coefficient as an optimization problem. The goal is to find $M + 1$ tuple $\mathbf{k} = [k_g, k_1, k_2, \dots, k_{M-1}, k_{ip}]$ that minimizes a certain reasonably chosen objective function and in the process produces a reasonable VT area profile that can be easily attained by the human vocal tract and produce the desired sound.

One of the first requirements is that the new lattice should have the same impulse response $\widehat{A}(z) = A^{(M)}(z)|_{\mathbf{k}}$ as that of M^{th} order linear predictor $A^{(M)}(z)$, so the most important objectives is to minimize \mathcal{J} .

$$\min_{-1 \leq \mathbf{k} \leq 1} 0.5 \sum_{i=1}^{M+1} \mathcal{J}_i \quad (8)$$

where \mathcal{J} is defined as the squared difference between the current and the target response of the lattice structure as shown in Figure 1.

$$\mathcal{J}_i = (\widehat{A}_i(z) - A_i(z))^2 \quad (9)$$

$\widehat{A}(z)$ and $A(z)$ are evaluated using Equation (7) with $k_g \neq 1$ and $k_g = 1$ respectively. Reflection coefficient k 's can be obtained by solving Equation (9) using any suitable non-linear optimization technique. Objective \mathcal{J} tries to ensure that the generated lattice has the same resonances (both frequencies and bandwidths) as ones produced by the traditional lattices.

In a strict signal processing sense, this lattice would have an identical response to the traditional lattice. On the other hand \mathcal{J} does not take into account the anatomical constraints imposed on articulators and does not guarantee producing a VT profile that can be attained by a human VT.

Some anatomical constraints have to be imposed on the objective, so the VT areas obtained by minimizing it have physical significance.

- *Constraint 1:* There should only be small changes in the areas of adjacent tubes.
- *Constraint 2:* The areas of the tubes in consecutive 10 ~ 30 ms frames of speech should be similar. The VT area profile across consecutive frames should not vary rapidly.

These constraints essentially imply that the VT has *finite flexibility* and an individual *cannot talk arbitrarily fast*. The optimization objective should also account for these constraints.

We select *Constraint 1* (Q) along with objective \mathcal{J} to form a new objective function. The new goal is to smooth the VT area profile and minimize the error in the lattice impulse response. The new optimization problem is stated in Equation (10)

$$\min_{-1 \leq \mathbf{k} \leq 1} 0.5 \sum_{i=1}^{M+1} (\mathcal{J}_i + Q_i) \quad (10)$$

where \mathcal{J} is defined by Equation (9) and Q is defines as

$$Q_i = (S_{(i-1)} - S_{(i)})^2 \quad (11)$$

The $M + 1$ tuple \mathbf{k} produced by minimizing the objective Equation (10) may not have $k_g = 1$, but these new sets of k 's still produce the same LPC polynomial $A(z)$. Physically the VT area profiles make more sense than the ones produced by using just objective \mathcal{J} .

4. Results

The results presented in this section address two aspects of the algorithm. The first group of results tries to evaluate the two objectives \mathcal{J} and $\mathcal{J} + Q$, the second group tries to verify if new set of VT areas adhere to the anatomical constraints.

The audio data used for the first set of results was a noise free, sustained vowel sound /a/. The audio was sampled at 16 kHz. 20 ms windows with 50% overlap were used. Hamming windows were applied to each frame before processing. No pre-emphasis was performed. A 15th order predictor was used giving us a 15 stage lattice/lossless tube model. A constrained optimization algorithm of Coleman et al. [9] was used to solve problems (8) and (10).

Figures 3(a) and 3(b) show the two VT area profiles obtained by minimizing \mathcal{J} . As seen in Figure 3(b) the second area profile can easily be attained by a human vocal tract but it is impossible for a vocal tract to attain the first shape 3(a). This was the major drawback in using only \mathcal{J} as the objective to be minimized. The objective \mathcal{J} does not impose any constraint on the areas of the tubes in space or time and may end up producing VT area profiles that are very abnormal.

Figures 3(a) and 3(c) show the area functions obtained using \mathcal{J} and $\mathcal{J} + Q$ as the objectives respectively, both of these area profiles were obtained for the same frame of speech, using the same starting point for the gradient descent algorithm. The orientation in Figure 3(c) can easily be achieved by the VT where as it is not possible to align the VT in the orientation shown in Figure 3(a). Minimizing objective (10) is like searching for an area-function that exhibits *maximal spatial smoothness* in the solution space.

For normal conversational speech, extracted VT areas profiles for neighboring frames should be very similar as indicated by *Constraint 2*. The Figure 4 shows area functions for 12

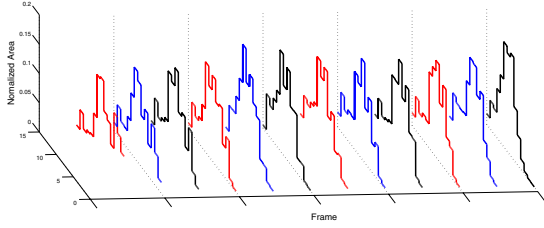


Figure 4: Area functions for 12 frames of speech showing the variation in the VT shape with time

consecutive frames of speech obtained by using Equation (10) as the objective. Changes in the areas of tubes in consecutive frames is very subtle making them very predictable. Next we present results comparing the predictability of the VT area profiles obtained using the gradient based and traditional approaches.

The N^{th} order model attempts to predict the area vector $\underline{S}_t = [S_1, S_2, \dots, S_M]^T$ for the t^{th} frame, as a linear combination of the N previous frames. Notationally we can write this as

$$\widehat{\underline{S}}_t = W_1 \underline{S}_{t-1} + W_2 \underline{S}_{t-2} + \dots + W_N \underline{S}_{t-N} \quad (12)$$

where $\widehat{\underline{S}}_t$ is the predicted value for the area vector \underline{S}_t , N is the order of the predictor and W_i 's are the predictor coefficient matrices. W_i 's are estimated by minimizing $\|\widehat{\underline{S}}_t - \underline{S}_t\|^2$, the squared error between the true and the predicted value. The matrix representation of Equation (12) is given by

$$\widehat{\underline{S}}_t = \mathbf{W} \mathbf{S}_{t-1} \quad (13)$$

Where $\mathbf{W} = [W_1, W_2, \dots, W_N]$ is the prediction matrix and $\mathbf{S}_t = [\underline{S}_t^T, \underline{S}_{t-1}^T, \dots, \underline{S}_{t-(N-1)}^T]^T$ is the vertical concatenation of area vectors for N previous frames. Minimization of the prediction error norm results in

$$\mathbf{W} = \underline{S}_t \cdot \mathbf{S}_{t-1}^\dagger \quad (14)$$

Where \dagger indicates the pseudo-inverse of the matrix.

Two sets of prediction matrices were trained, one for area profiles obtained using the traditional approach (Itakura-Saito) and the second for areas obtained using gradient based method (Objective $\mathcal{J} + \mathcal{Q}$). Simulations were performed using Wall Street Journal audio database; half the dataset was used for training \mathbf{W} . These \mathbf{W} 's were used for calculating the VT area prediction errors on other half of the data. 20 ms frames with 50% overlap were used. No pre-emphasis was performed. Hamming windows were applied to each frame. Prediction matrices (\mathbf{W}) up to 10^{th} order were trained.

Figure 5 shows the mean squared prediction error for both the methods of VT area estimation. For both the methods, estimation error decreases with the increase in prediction order. As seen from the plots there is almost 38% reduction in prediction error when the VT areas are estimated from k 's obtained by minimizing the objective (10).

5. Conclusion

In this paper we presented a framework based on the lattice representation of VT for estimation of the glottal and lip reflection coefficients from only speech data. PARCOR's estimation is formulated as an optimization problem. One of the objectives for optimization is to match the impulse response of the lattice

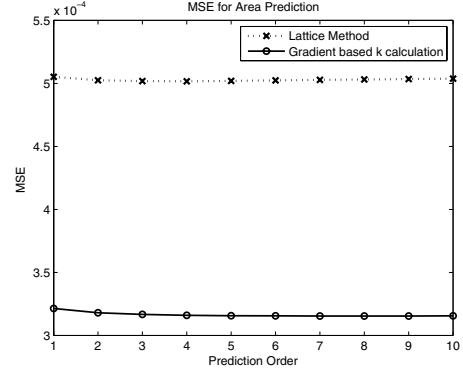


Figure 5: Area Function Prediction Error

to that of a linear prediction polynomial of the same order obtained by using Levinson's recursion.

The algorithm does not make assumptions on losses in the VT. The framework allows for the distribution of the losses in the VT, at both the lip and the glottal end.

We demonstrated that by selecting a suitable objective for minimization, it is possible to obtain VT area profiles that are smoother than those obtained traditional lattice solutions. This method also ensures reduction in the variation of VT area across frames thereby making them very predictable. This predictability can be exploited for formant tracking, de-noising and compression of speech, which are currently under investigation.

6. References

- [1] G Fant, *Acoustic Theory of Speech Production*, Mouton, Le Hague, 1960.
- [2] A. R. Greenwood, C. C. Goodyear, and P. A. Martin, "Measurements of vocal tract shapes using magnetic resonance imaging," *Comm., Speech and Vision, IEE Proc I*, vol. 139(6), pp. 219–226, December 1992.
- [3] B. H. Story, Titze I. R., and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *JASA*, vol. 1, pp. 537–554, July 1996.
- [4] B Gopinath and M. M. Sondhi, "Determination of shape of the human vocal tract by acoustic measurements," *Bell Sys. Tech. Journal*, vol. 49, pp. 1195–1214, 1970.
- [5] H Deng, R.K. Ward, M.P. Beddoes, and D. O'Shaughnessy, "Obtaining lip and glottal reflection coefficients from vowel sounds," *ICASSP*, vol. 1, pp. 1–373–376, May 2006.
- [6] B. S. Atal and Hanauer S. L., "Speech analysis and synthesis by linear prediction of speech wave," *JASA*, vol. 1, pp. 637–655, August 1971.
- [7] H Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveform," *IEEE Trans. on Audio and Electroacoustics*, vol. 21(5), pp. 417–427, October 1973.
- [8] M. M Sondhi, "Estimation of vocal tract areas: the need for acoustical measurements," *IEEE Trans. on Acoustic Speech and Signal Processing*, vol. 27(3), pp. 268–273, June 1979.
- [9] Coleman T. F. and Yuying Li, "An interior trust region approach for nonlinear minimization subject to bounds," *SIAM Journal on Optimization*, vol. 6(2), pp. 418–445, 1996.