IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Acoustico-articulatory inversion of unequal-length tube models through lattice inverse filtering

Sacha KRSTULOVIĆ

IDIAP–RR 98-16

December 1998

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP Research Report 98-16

Acoustico-articulatory inversion of unequal-length tube models through lattice inverse filtering

Sacha KRSTULOVIĆ

December 1998

Abstract. Constraints related to the Distinctive Regions and Modes (DRM) speech production model are incorporated in the framework of speech analysis by inverse filtering. It is shown that the analogy between Auto-Regressive modeling and acoustic models based on acoustic tubes is still respected when using tubes with unequal length elementary sections. This analogy can be exploited for the inversion of the synthesis process using a method related to Burg's estimation method. Experimental results show that an improvement over traditional methods is brought by the DRM-related constraints included in the parameter estimation scheme.

Acknowledgements: This work is supported by Swiss National Science Fund grant nr. 2100-49'725.96 for the *ARTIST* project. The author wishes to thank Dr. Chafic Mokbel for many fruitful discussions, and Dr. Gerard Chollet for having been the initiator of the project.

Contents

1 Introduction

Traditional speech analysis methods are based on general purpose signal processing algorithms. They do not take advantage of the particular nature of speech acoustics. Hence, a better adaptation of signal processing methods to speech acoustics may help reaching better speech modeling performances.

The adaptation of analysis methods can be performed with the help of constraints related to speech production and applied in the course of the analysis protocol. This supposes that a signal processing algorithm and a speech production model are chosen, and a link between the two is established.

The present study exposes a link between Auto-Regressive (AR) modeling, also known as Linear Prediction modeling, and the Distinctive Regions and Modes (DRM) speech production model. This link is exploited to build a speech analysis method that incorporates the DRM constraints.

Section 2 will present the background, objectives and paradigm of the DRM model. Section 3 will explain in detail how to relate the DRM model to a linear prediction model. Starting from fluid dynamics equations, the derivation of recursive formulas for the computation of the transfer function of an acoustic tube with elementary sections of any lengths will be exposed. The stability of the obtained transfer function will be questioned. Lattice forms for the implementation of the transfer function will be described. In section 4, a method to estimate the transfer function's parameters from the speech waveform will be exposed. This method is inspired from the well known Burg method [Mak77]. It allows to recover the shape of the described acoustic tubes, which amounts to inverting the synthesis process outlined in section 3. Section 5 will describe the experiments aiming at assessing the performances of the inversion process, together with their results.

2 The DRM model

2.1 Background and objectives

The emergence of the DRM model is related to the problem of modeling the relationship between human articulation and its acoustic effects. Fant [Fan73] and Fant and Pauli [FP74] studied the evolution of vocal tract resonances in relation with its shape. Starting from their work, M. Mrayati, R. Carré and B. Guérin have proposed a new model [MCG88]. Their goal was to provide a simple model that would allow to "pilot" the formant trajectories from a deformable acoustic tube. The tube would be made of 8 connected cylindrical sections of fixed lengths and of varying section, and it would be excited by a glottal waveform source or a noise source. This tube, although accounting for constraints related to the vocal tract physiology, didn't pretend to be an exact modeling of the vocal tract shape, but more a model of the vocal tract behavior.

The study of Mrayati, Carré and Guérin has led to the DRM paradigm, which is described hereafter.

2.2 The DRM paradigm: configurations, modes, regions

2.2.1 Origin of the model

Fant and Pauli have represented the relationship between the evolution of formant frequencies and small variations of areas in an acoustic tube (closed at one end and open at the other) using a function called "sensitivity function". This function involves the potential energy and the kinetic energy of a considered waveform. It is defined as:

$$
\begin{aligned}
\frac{\Delta F_i}{F_i} &= \sum_{n=1}^{N} S_n \frac{\Delta A_n}{A_n} \\
&= \sum_{n=1}^{N} \frac{\bar{E}c_n - \bar{E}p_n}{\bar{E}tot_n} \frac{\Delta A_n}{A_n}
\end{aligned}
$$

with: F_i : frequency of the ith formant,
n : sections' indices,
S_n : value of the sensitivity function at section n,
A_n : area of section n.

The sensitivity function involves air velocity, air pressure and areas of the tube, since:

$$\bar{E}c = f(\vec{v}(x)^2, \frac{1}{A(x)}), \quad \text{where } \vec{v}(x) \text{ is air velocity.}$$

$$\bar{E}p = f(\vec{P}(x)^2, A(x)), \quad \text{where } \vec{P}(x) \text{ is air pressure.}$$

A positive value of the sensitivity function for a given section and a given formant means that increasing the section's area will increase the formant frequency. Conversely, a negative value of the sensitivity function will indicate an inverse variation of the areas and frequencies (increased area $\Longleftrightarrow$ decreased frequency).

Mrayati, Carré and Guérin have observed that the zero crossings of the sensitivity function were stable for certain ranges of sections' areas. To classify these ranges, they defined two tube *configurations*: the closed-open tube, where glottis end is closed and lips end is open, and the closed-closed tube, where the lips end is closed. They further divided the closed-open configuration in two *modes*, where the sensitivity function had different (stable) zero-crossing locations: the One Tract Mode (OTM), where the tube presents no very narrow constrictions, and the Two Tracts Mode (TTM), where the tube presents one narrow constriction.

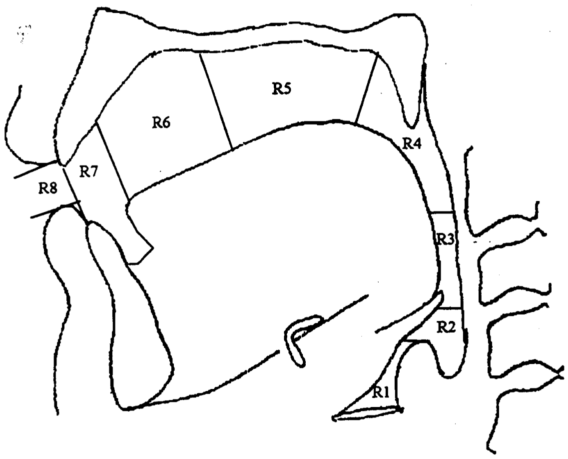
They have thereafter defined eight distinctive tube regions, delimited by the stable zero crossings of the sensitivity function. Inside these regions, the behavior of the formants as a function of the area variations is stable and monotonic for specific area ranges.

This phenomenon is described in more detail for the OTM in the next section.

2.2.2 One Tract Mode (OTM)

The One Tract Mode (OTM) corresponds to a closed-open tube configuration where no narrow constriction exists. From an acoustic point of view, there is an important coupling between the front and the back of the tube (hence the name of the mode). This mode is defined precisely in figure 1 and table 1.

Several useful remarks can be formulated. The first is that the distinctive regions roughly correspond to the position of human articulators:



(From [Che95])

The second is that all possible combinations of formant frequencies increase/decrease versus sections' area increase/decrease can be observed: the model is said to be "pseudo-orthogonal". This characteristic of the OTM can be easily verified from table 1.

2.2.3 Other configurations and modes

As explained earlier, the taxonomy of the DRM model's configurations and modes comprises also:

- the Two Tracts Mode (TTM) when the tube comprises a very narrow constriction, and thus acoustic coupling between front and back is low
- the Transition Mode (TM), between the OTM and the TTM
- the Closed-Closed Tube configuration mode (CCTM) when the lips are closed.

These modes define a different repartition for the distinctive regions and a different area/acoustic variations relationship. **However, in the present study, we will consider that speech is produced mainly in the OTM mode.** This approximation is reasonable since the time intervals during which the tube reaches CCTM (front plosives) or TTM (middle and back plosives) are usually very short. Consequently, the repartition of the distinctive regions will be fixed to that of the OTM, as given in table 1 and considering that the total tube length L is 17cm (which is an average for adult speakers).

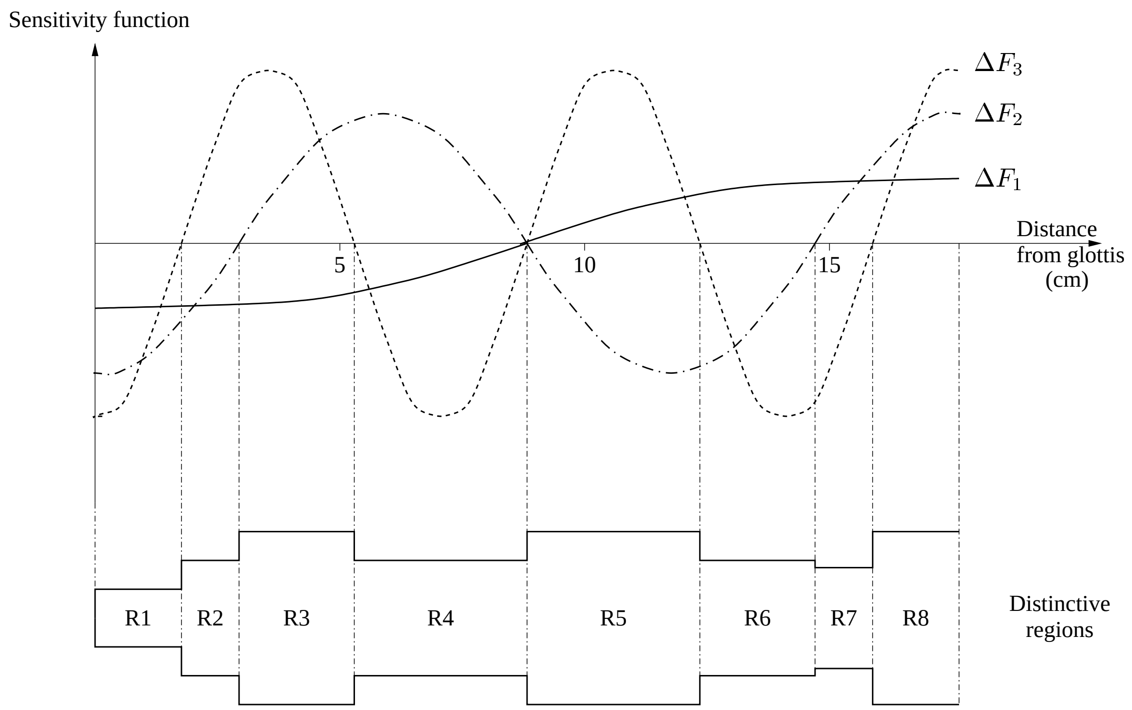


Figure 1: Relation between sensitivity functions and distinctive areas in the One Tract Mode (adapted from [Che95]).

		R1	R2	R3	R4	R5	R6	R7	R8
Section lengths		$L/10$	$L/15$	$2L/15$	$L/5$	$L/5$	$2L/15$	$L/15$	$L/10$
Area ranges	**from**	0.075	2.5	1.2	2.5	0.43	1.9	2	0.02
(cm^2)	**to**	15	15	15	15	5.3	15	5.3	15
Formant behavior	**F1**	- $\searrow$	+ $\nearrow$	+ $\nearrow$	- $\searrow$	+ $\nearrow$	- $\searrow$	- $\searrow$	+ $\nearrow$
for an *increasing* area	**F2**	- $\searrow$	- $\searrow$	+ $\nearrow$	+ $\nearrow$	- $\searrow$	- $\searrow$	+ $\nearrow$	+ $\nearrow$
(area $\nearrow$)	**F3**	- $\searrow$	- $\searrow$	- $\searrow$	- $\searrow$	+ $\nearrow$	+ $\nearrow$	+ $\nearrow$	+ $\nearrow$

L is the total tube length; + or - indicate the sign of the sensitivity function.

Table 1: Relation between area variations and formant frequency variations in the One Tract Mode.

2.3 Known model limitations

The DRM model is known to have the following limitations :

- it doesn't include the nasal tract, thus being inaccurate for the modeling of nasal vowels.
- from a physiological point of view, it is a rather crude articulatory model.
- it doesn't include the production of unvoiced plosives and fricatives.

Despite these limitations, it remains a good candidate to model the acoustic production of "most of speech".

3 Generalized filtering process of an acoustic tube

3.1 Development of the transfer function

3.1.1 Fluid dynamics basis of the problem

Basic system :

The vocal tract is considered to be an acoustic tube discretized in M cylindrical sections of *various lengths*. The cross-sectional areas of the sections are allowed to vary across time. The lengths of the section remain fixed. Sections are numbered in crescent order *from lips to glottis*.

Assumptions :

- sound waves are plane fluid waves ([Fla72], pp.24-25; [MI68], p.467).
- the tube is rigid (no wall impedance is considered).
- tube quantization introduces acceptable error if lengths of the sections are kept short compared to a wavelength at the highest frequency of interest ([Fla72], p.25) :

  $$L_{max} \ll \frac{c}{F_{max}}$$

  where c is speed of sound.
- energy losses due to viscosity of air and heat conduction are neglected.

Equation set :

Posing the problem in terms of fluid dynamics, we can consider that the volume velocity $u_m(t,d)$ and the pressure $p_m(t,d)$ in section m derive from a potential $\Phi_m(t,d)$:

$$u_m(t,d) = -S_m \frac{\partial \Phi_m(t,d)}{\partial d} \tag{1}$$

$$p_m(t,d) = \rho \frac{\partial \Phi_m(t,d)}{\partial t} \tag{2}$$

with :

- t : time variable
- d : distance variable
- S_m : cross-sectional area of m^{th} section
- ρ : density of air

The evolution of the fluid state is thereafter described by Webster's equation [Bon83] :

$$\frac{\partial^2 \Phi_m(t,d)}{\partial d^2} - \frac{1}{c^2} \frac{\partial^2 \Phi_m(t,d)}{\partial t^2} = 0 \tag{3}$$

where c denotes the sound velocity.

Equation solving :

If we assume that the excitation source (the "glottis" of the tube) delivers a sinusoidal signal, then the solution of this differential equation is of the classic form :

$$\Phi_m(t,d) = A \exp^{j\omega(t-d/c)} + B \exp^{j\omega(t+d/c)} \tag{4}$$

where A and B are constants[1]. Remarking that $u_m(t,d)$ can be decomposed into a forward-traveling wave $u_m^+(t,d)$ and a backward-traveling wave $u_m^-(t,d)$, the above solution can be decomposed in the following way :

$$\begin{cases} u_m(t,d) = u_m^+(t,d) - u_m^-(t,d) \\ p_m(t,d) = \frac{\rho c}{S_m}\left\{u_m^+(t,d) + u_m^-(t,d)\right\} \end{cases} \tag{5}$$

with

$$\begin{cases} u_m^+(t,d) = \frac{j\omega S_m A}{c} \exp^{j\omega(t-d/c)} \\ u_m^-(t,d) = \frac{j\omega S_m B}{c} \exp^{j\omega(t+d/c)} \end{cases} \tag{6}$$

At the connection between section m and section $m+1$, the volume velocity and pressure must be continuous. We therefore have the additional relations :

$$\begin{cases} u_{m+1}(t,d_m) = u_m(t,d_m) \\ p_{m+1}(t,d_m) = p_m(t,d_m) \end{cases} \tag{7}$$

with d_m being the distance between the glottis and the connection of sections m and $m+1$ (see figure 2). Since the speed of sound is constant, the distance variable can be related to the time variable and can thus be eliminated. Since there is no loss in a particular section, we also have, *inside* the limits of a section :

$$\begin{cases} u_m^+(t,d) = u_m^+(t,d-\Delta l_m) = u_m^+(t - \frac{\Delta l_m}{c}) \\ u_m^-(t,d) = u_m^-(t,d-\Delta l_m) = u_m^-(t + \frac{\Delta l_m}{c}) \end{cases} \tag{8}$$

with Δl_m being the length of the considered piece of tube.

This step is very important, since dropping the distance variable allows us to express our problem in terms of time series analysis. Furthermore, the fact that the problem can be solved considering only the information available at tubes' junctions will allow us to work in a discrete signal processing framework.

3.1.2 From fluids to signals

Starting from the solution of fluid dynamics equations, and injecting (5) and (8) in (7), we end up with the following relations between the forward and backward traveling waves at each junction :

$$\begin{cases} u_{m+1}^+(t-\Delta_m t) - u_{m+1}^-(t+\Delta_{m+1} t) = u_m^+(t) - u_m^-(t) \\ \frac{\rho c}{S_{m+1}}\left\{u_{m+1}^+(t-\Delta_m t) + u_{m+1}^-(t+\Delta_{m+1} t)\right\} = \frac{\rho c}{S_m}\left\{u_m^+(t) + u_m^-(t)\right\} \end{cases} \tag{9}$$

where :

$$\Delta_m t = \frac{\Delta l_m}{c}$$

is the time necessary for a wave to travel through a tube of length Δl_m.

[1] If the excitation signal is made of a linear combination of sine waves, which is the case for signals admitting a Fourier Series development, the corresponding solution is a linear combination of the solutions for any individual sinusoidal component. In the case of speech, the relations developed here do not loose their generality (in the limits of the assumptions made at the beginning) since the non-sinusoidal excitation such as the wave coming out of the vocal cords admits a development in Fourier series.

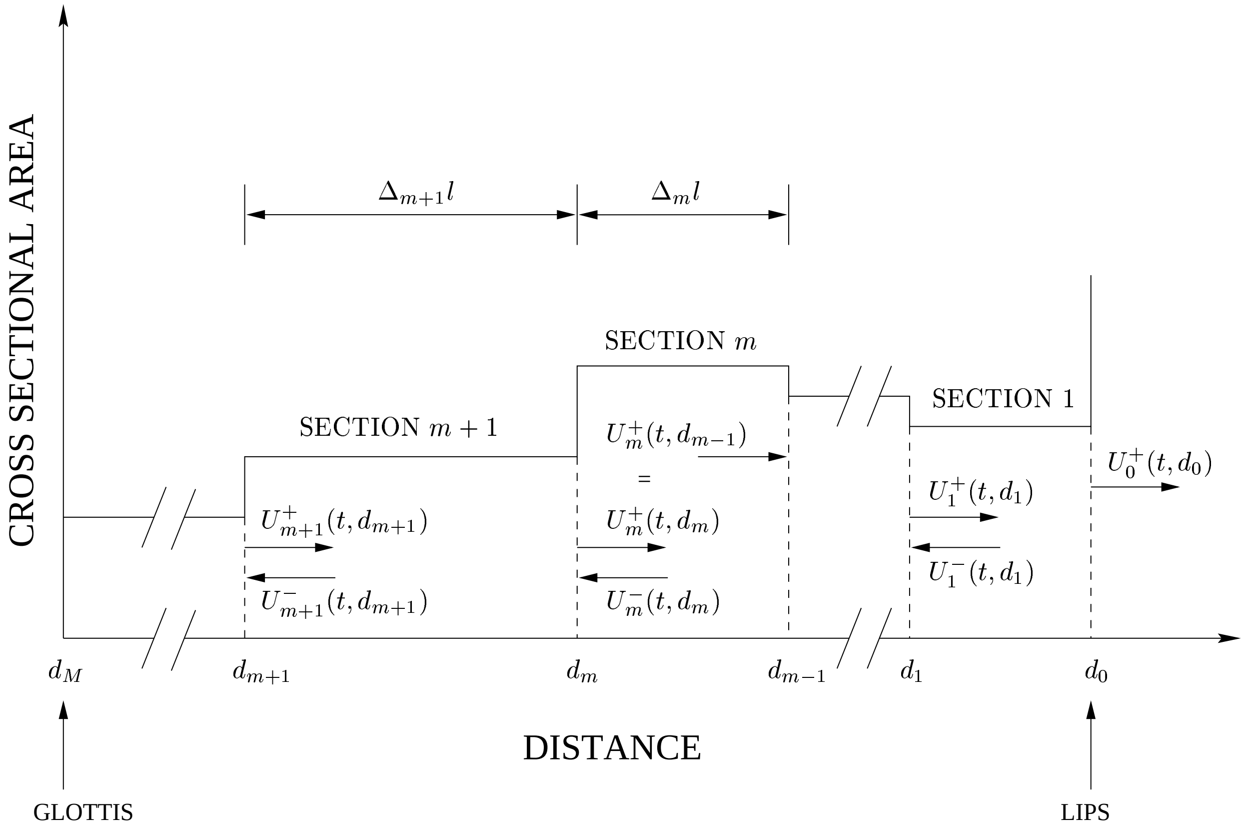


Figure 2: Non-uniform acoustic tube model of the vocal tract. (Adapted from [Wak73].)

Now defining reflection coefficients as :

$$\mu_m = \frac{S_{m+1} - S_m}{S_{m+1} + S_m}$$

and applying to the above equations, we obtain :

$$\left\{ \begin{array}{l} u^+_{m+1}(t - \Delta_m t) = \frac{1}{1-\mu_m} \left\{ u^+_m(t) + \mu_m u^-_m(t) \right\} \\ \\ u^-_{m+1}(t + \Delta_m t) = \frac{1}{1-\mu_m} \left\{ \mu_m u^+_m(t) + u^-_m(t) \right\} \end{array} \right. \tag{10}$$

Defining a unit length Δl_{unit} as the greatest common divisor of the lengths Δl_m, we can apply the z-transform with z defined as $z = e^{j2\omega\Delta l_{unit}/c} = e^{j2\omega\Delta_{unit}t}$, and we obtain :

$$\left\{ \begin{array}{l} z^{\frac{-n_m}{2}} U^+_{m+1}(z) = \frac{1}{1-\mu_m} \left[U^+_m(z) + \mu_m U^-_m(z) \right] \\ \\ z^{\frac{n_m}{2}} U^-_{m+1}(z) = \frac{1}{1-\mu_m} \left[\mu_m U^+_m(z) + U^-_m(z) \right] \end{array} \right. \tag{11}$$

i.e.

$$\left\{ \begin{array}{l} U^+_{m+1}(z) = \frac{z^{\frac{n_m}{2}}}{1-\mu_m} \left[U^+_m(z) + \mu_m U^-_m(z) \right] \\ \\ U^-_{m+1}(z) = \frac{z^{\frac{-n_m}{2}}}{1-\mu_m} \left[\mu_m U^+_m(z) + U^-_m(z) \right] \end{array} \right. \tag{12}$$

and, in matrix notation :

$$\left[\begin{array}{c} U^+_{m+1}(z) \\ U^-_{m+1}(z) \end{array} \right] = \frac{z^{\frac{n_m}{2}}}{1-\mu_m} \left[\begin{array}{cc} 1 & \mu_m \\ \mu_m z^{-n_m} & z^{-n_m} \end{array} \right] \left[\begin{array}{c} U^+_m(z) \\ U^-_m(z) \end{array} \right] \tag{13}$$

3.1.3 Appearance of the transfer function

If we assume that the section corresponding to the lips extremity is connected to a tube of infinite section, it amounts to the following boundary condition at front end (or lips end) of our model :

$$S_{-1} = \infty \;\Rightarrow\; \mu_0 = -1$$

Applying this condition, we can write :

$$\left[\begin{array}{c} U^+_{m+1}(z) \\ U^-_{m+1}(z) \end{array} \right] = z^{\frac{1}{2}\sum_{k=0}^{m} n_k} \, K_m \left[\begin{array}{c} D^+_m(z) \\ D^-_m(z) \end{array} \right] \left\{ U^+_0(z) - U^-_0(z) \right\} \tag{14}$$

with

$$\left[\begin{array}{c} D^+_m(z) \\ D^-_m(z) \end{array} \right] = \left[\begin{array}{cc} 1 & \mu_m \\ \mu_m z^{-n_m} & z^{-n_m} \end{array} \right] \left[\begin{array}{cc} 1 & \mu_{m-1} \\ \mu_{m-1} z^{-n_{m-1}} & z^{-n_{m-1}} \end{array} \right] \cdots \left[\begin{array}{cc} 1 & \mu_1 \\ \mu_1 z^{-n_1} & z^{-n_1} \end{array} \right] \left[\begin{array}{c} 1 \\ -z^{-n_0} \end{array} \right] \tag{15}$$

and

$$K_m = \prod_{i=0}^{m} \frac{1}{1-\mu_i} \tag{16}$$

Neglecting the overall delay $z^{\frac{1}{2}\sum_{k=0}^{m} n_k}$ and the gain K_m, **the true transfer function for the volume velocity when travelling from lips to glottis is there denoted by $D_m^+(z)$, and can be built recursively by applying :**

$$
\begin{aligned}
\begin{bmatrix} D_{m+1}^+(z) \\ D_{m+1}^-(z) \end{bmatrix} &= \begin{bmatrix} 1 & \mu_{m+1} \\ \mu_{m+1} z^{-n_{m+1}} & z^{-n_{m+1}} \end{bmatrix} \begin{bmatrix} D_m^+(z) \\ D_m^-(z) \end{bmatrix} \\
\begin{bmatrix} D_0^+(z) \\ D_0^-(z) \end{bmatrix} &= \begin{bmatrix} 1 \\ -z^{-n_0} \end{bmatrix}
\end{aligned} \tag{17}
$$

What is described by this recursion is the growth of the transfer function corresponding to the inverse of the vocal tract filtering action : as indicated, it concerns a wave that would travel from lips to glottis after an excitation at lips end. This recursion is of interest in the current framework, since the goal is to estimate the model's parameters through inverse filtering of the produced speech waveform.

The glottis-to-lips ("synthesis oriented") transfer function can be deduced very simply from $D_m^+(z)$ since it corresponds to its inverse : $A_m^+(z) = 1/D_m^+(z)$. The recursion giving directly the growth of $A_m^+(z)$ can be easily derived by inverting relations (12) and expressing the result in matrix form.

3.1.4 Additional relations

We can show by mathematical induction from equation (15) that we have :

$$D_m^-(z) = -z^{-\sum_{k=0}^{m} n_k} D_m^+(1/z) \tag{18}$$

This means that the forward inverse filtering function $D_m^+(z)$ is the time shifted reciprocal of the backward inverse filtering function $D_m^-(z)$.

Developing (17) and applying (18), we obtain :

$$
\begin{cases}
D_{m+1}^+(z) = D_m^+(z) - \mu_{m+1}\ z^{-\sum_{k=0}^{m} n_k}\ D_m^+(1/z) \\[2ex]
D_{m+1}^+(1/z) = -\mu_{m+1}\ z^{\sum_{k=0}^{m} n_k}\ D_m^+(z) + D_m^+(1/z)
\end{cases} \tag{19}
$$

We can remark that if we change the variable z to $1/z$ in the first of the above formulae, we obtain the formula in the second line. Both formulae are equivalent with regard to the relationship they imply between $D_{m+1}^+(z)$ and $D_m^+(z)$. We could use this recursion to build the forward transfer function $D_m^+(z)$ (starting from $D_0^+(z) = 1$) instead of the matricial recursion (17).

We will now develop these relations in order to study more precisely the form and the growth of the inverse filtering transfer function.

3.2 Auto-Regressive (AR) nature of the transfer function

The development is made in the case of equal length sections and then in the case of unequal-length sections as illustrated by the DRM case.

3.2.1 Classical case : the length of the sections is uniform

In this case, the transmission delay induced in every piece of tube is the same. Therefore, we can set $n_k = 1 \quad \forall k$, i.e. $z^{-n_k} = z^{-1} \quad \forall k$ in all the above equations. z is then defined as $z = e^{j\omega 2\Delta l/c}$, Δl being the length of every cylindrical piece of tube.

We know from equation (15) that $D_m^+(z)$ is of the form :

$$D_m^+(z) = \sum_{i=0}^{m} a_i^{(m)} z^{-i} \tag{20}$$

which corresponds to a Finite Impulse Response (FIR) inverse filter. When used for acoustic filtering in the "synthesis direction" (from glottis to lips), it therefore acts as an Infinite Impulse Response (IIR) filter. Hence, in the synthesis framework, speech production is considered as an Auto-Regressive (AR) process [MG76].

When adding a new cylindrical section to our tube, we have :

$$D_{m+1}^+(z) = \sum_{i=0}^{m+1} a_i^{(m+1)} z^{-i} \tag{21}$$

which means that the degree of the polynomial inverse transfer function increases by one at each tube growing step.

Now, we can try and formalize the relation between the transfer function coefficients a_k and the reflection coefficients μ_i. The relation (19) gives :

$$\sum_{i=0}^{m+1} a_i^{(m+1)} z^{-i} = \sum_{i=0}^{m} a_i^{(m)} z^{-i} - \mu_{m+1} z^{-(m+1)} \sum_{i=0}^{m} a_i^{(m)} z^{i} \tag{22}$$

i.e.

$$\sum_{i=0}^{m+1} a_i^{(m+1)} z^{-i} = \sum_{i=0}^{m} a_i^{(m)} z^{-i} - \mu_{m+1} \sum_{i=0}^{m} a_i^{(m)} z^{i-(m+1)} \tag{23}$$

or, changing the mute index i to $(m+1-i)$ in the second sum of the right term :

$$\sum_{i=0}^{m+1} a_i^{(m+1)} z^{-i} = \sum_{i=0}^{m} a_i^{(m)} z^{-i} - \mu_{m+1} \sum_{i=1}^{m+1} a_{m+1-i}^{(m)} z^{-i} \tag{24}$$

Identifying the coefficients of the polynoms in z^{-i} on each side of the equal sign, we obtain :

$$\begin{cases} a_0^{(m+1)} = a_0^{(m)} \\ a_1^{(m+1)} = a_1^{(m)} - \mu_{m+1} a_m^{(m)} \\ \vdots \\ a_m^{(m+1)} = a_m^{(m)} - \mu_{m+1} a_1^{(m)} \\ a_{m+1}^{(m+1)} = -\mu_{m+1} a_0^{(m)} \end{cases} \tag{25}$$

which can be formalized in a simple form as :

$$\begin{cases} a_i^{(m+1)} = a_i^{(m)} - \mu_{m+1} a_{m+1-i}^{(m)} \\ a_0^{(0)} = 1 \\ a_k^{(0)} = 0, \quad \forall k \neq 0 \end{cases} \tag{26}$$

These equations are similar to those linking partial correlation coefficients μ_i to prediction coefficients a_i in the Levinson-Durbin algorithm for LPC modeling. If we analyze speech with a sampling frequency of $F_s = \frac{c}{2\Delta l}$, and if we can estimate the reflection coefficients in a reliable way, the equivalence between an Auto-Regressive model of order M and the filtering process of an evenly discretized tube needs no more assumptions to hold.

3.2.2 A more general case : the length of the sections is not uniform

If the length of the sections is not uniform, such as in the DRM, we must deal with the irregular delays and the z^{-n_k} not being equal to z^{-1}. To formalize the growth of the transfer function in a readable way, we will borrow the notation of the summation indexes to the set theory. Let Ω_m be the set of all possible indexes k for the discrete delays n_k met in a particular tube, and let Γ be a set containing one of the possible index combinations[2].

We know from equation (15) that $D_m^+(z)$ is a polynomial in z and we can now express its form as :

$$D_m^+(z) = \sum_{\Gamma \subset \{0,1,\cdots m\}} a_\Gamma^{(m)} z^{-\sum_{k\in\Gamma} n_k} \tag{27}$$

This transfer function has a special (constrained) form : it still corresponds to an IIR filter (still FIR/AR when reverted for synthesis), but not all the polynomial degrees are represented. As before, we can observe the growth of the inverse transfer function :

$$D_{m+1}^+(z) = \sum_{\Gamma \subset \{0,1,\cdots m+1\}} a_\Gamma^{(m+1)} z^{-\sum_{k\in\Gamma} n_k} \tag{28}$$

We see here that when going from step (m) to step $(m+1)$, the polynomial degree increases by $-n_{m+1}$, and that not all the degrees between 0 and $-\sum_{k\in\Gamma} n_k$ are guaranteed to be represented.

Coming back to the development of the $a_k \Leftrightarrow \mu_i$ relation, equation (19) now gives :

$$\sum_{\Gamma\subset\Omega_{m+1}} a_\Gamma^{(m+1)} z^{-\sum_{k\in\Gamma} n_k} = \sum_{\Gamma\subset\Omega_m} a_\Gamma^{(m)} z^{-\sum_{k\in\Gamma} n_k} - \mu_{m+1} z^{-\sum_{k\in\Omega_{m+1}} n_k} \sum_{\Gamma\subset\Omega_m} a_\Gamma^{(m)} z^{\sum_{k\in\Gamma} n_k} \tag{29}$$

i.e.

$$\sum_{\Gamma\subset\Omega_{m+1}} a_\Gamma^{(m+1)} z^{-\sum_{k\in\Gamma} n_k} = \sum_{\Gamma\subset\Omega_m} a_\Gamma^{(m)} z^{-\sum_{k\in\Gamma} n_k} - \mu_{m+1} \sum_{\Gamma\subset\Omega_m} a_\Gamma^{(m)} z^{\sum_{k\in\Gamma} n_k - \sum_{k\in\Omega_{m+1}} n_k} \tag{30}$$

For a particular subset Γ of our index set Ω_m, we can show the following :

$$\begin{aligned} \textstyle\sum_{k\in\Gamma} n_k - \sum_{k\in\Omega_{m+1}} n_k &= \underbrace{\sum_{k\in\Gamma} n_k - \sum_{k\in\Omega_m} n_k}_{} \quad -n_{m+1} \\ &= \quad -\textstyle\sum_{k\in\overline{\Gamma}} n_k \qquad -n_{m+1} \end{aligned} \tag{31}$$

$\overline{\Gamma}$ being the complementary set of Γ so that $\Gamma \cup \overline{\Gamma} = \Omega_m$. Equation (30) then becomes :

$$\sum_{\Gamma\subset\Omega_{m+1}} a_\Gamma^{(m+1)} z^{-\sum_{k\in\Gamma} n_k} = \sum_{\Gamma\subset\Omega_m} a_\Gamma^{(m)} z^{-\sum_{k\in\Gamma} n_k} - \mu_{m+1} \sum_{\Gamma\subset\Omega_m} a_\Gamma^{(m)} z^{-\sum_{k\in\overline{\Gamma}} n_k + n_{m+1}} \tag{32}$$

In this case, the analytical identification of the polynomial coefficients has to be performed on a case-by-case basis.

For instance, let us express it in the case of the DRM. In this case, we have 8 sections of unequal length with $\Delta l_{unit} = L/30$ (L being the total length of the full tube). The lengths of the sections are distributed as follows from lips to glottis : $\Delta l_0 = 3\Delta l_{unit}$, $\Delta l_1 = 2\Delta l_{unit}$, $\Delta l_2 = 4\Delta l_{unit}$, $\Delta l_3 = 6\Delta l_{unit}$, $\Delta l_4 = 6\Delta l_{unit}$, $\Delta l_5 = 4\Delta l_{unit}$, $\Delta l_6 = 2\Delta l_{unit}$, $\Delta l_7 = 3\Delta l_{unit}$. Hence, in the DRM case, recursion (17) expands in the matricial form :

[2]We have $\Omega_m = \{0, 1, \cdots m\}$ and $\Gamma \subset \Omega_m$. This means that Γ belongs to the set of all subsets of Ω_m. We can remark that this later set defines a σ-algebra on the set of delays n_k. A measure on this set could be defined as $\sum_{k\in\Gamma} n_k$. We don't know if such measure theory notions have already been used in the framework of polynomial transfer functions analysis, but researchers interested in measure theory might find here a lead to an alternate way of formalizing the problem.

$$\left[\begin{array}{c} D^+(z) \\ D^-(z) \end{array}\right] = \left[\begin{array}{cc} 1 & \mu_7 \\ \mu_7 z^{-3} & z^{-3} \end{array}\right] \left[\begin{array}{cc} 1 & \mu_6 \\ \mu_6 z^{-2} & z^{-2} \end{array}\right] \cdots \left[\begin{array}{cc} 1 & \mu_2 \\ \mu_2 z^{-4} & z^{-4} \end{array}\right] \left[\begin{array}{cc} 1 & \mu_1 \\ \mu_1 z^{-2} & z^{-2} \end{array}\right] \left[\begin{array}{c} 1 \\ -z^{-3} \end{array}\right] \tag{33}$$

When observing the growth of the inverse transfer function between, for instance, step 3 and step 4 (see equations developed in figure 4), and replacing the Γ indexes by integer indexes corresponding to the place of the increasing negative powers of z, we obtain the following set of equations :

$$\left\{\begin{array}{lcll} a_0^{(4)} & = & a_0^{(3)} & = 1 \\ a_1^{(4)} & = & a_1^{(3)} & \\ a_2^{(4)} & = & a_2^{(3)} & \\ a_3^{(4)} & = & a_3^{(3)} & \\ a_4^{(4)} & = & a_4^{(3)} & \\ a_5^{(4)} & = & a_5^{(3)} & -\mu_4 a_7^{(3)} \\ a_6^{(4)} & = & a_6^{(3)} & \\ a_7^{(4)} & = & & -\mu_4 a_6^{(3)} \\ a_8^{(4)} & = & a_7^{(3)} & -\mu_4 a_5^{(3)} \\ a_9^{(4)} & = & & -\mu_4 a_4^{(3)} \\ a_{10}^{(4)} & = & & -\mu_4 a_3^{(3)} \\ a_{11}^{(4)} & = & & -\mu_4 a_2^{(3)} \\ a_{12}^{(4)} & = & & -\mu_4 a_1^{(3)} \\ a_{13}^{(4)} & = & & -\mu_4 \end{array}\right. \tag{34}$$

Therefore, in the general case, we see that if we try to operate a polynomial coefficients identity starting from equation (19), we cannot meet the Levinson-Durbin relation tying prediction coefficients a_i and reflection coefficients μ_i.

Consequently, we cannot estimate the parameters of those inverse filters, corresponding to tubes with non-even section lengths, through direct solving of the Yule-Walker equations, as does the Levinson-Durbin algorithm.

3.2.3 Taking the problem by the other end : effect of $\mu_i = 0$ in a step of the Levinson recursion

An equivalent way of trying to solve our process identification problem was by starting with the Levinson recursion and disturbing it by adding the constraints inherited from the special structure of unequal-lengths tube models.

As seen above, we can consider that an unequal-lengths tube is made of a pile of equal-length elementary sections, some of them being fastened together. This amounts to setting some reflection coefficients to zero in the course of the Levinson-Durbin algorithm evaluation.

$$\begin{bmatrix} 1 \\ -z^{-1} \end{bmatrix}$$

$$\begin{bmatrix} 1-\frac{\mu_1}{z} \\ \frac{\mu_1}{z}-z^{-2} \end{bmatrix}$$

$$\begin{bmatrix} 1+\frac{\mu_2\mu_1-\mu_1}{z}-\frac{\mu_2}{z^2} \\ \frac{\mu_2}{z}+\frac{\mu_1-\mu_2\mu_1}{z^2}-z^{-3} \end{bmatrix}$$

$$\begin{bmatrix} 1+\frac{\mu_2\mu_1-\mu_1+\mu_3\mu_2}{z}+\frac{-\mu_2+\mu_3\mu_1-\mu_3\mu_2\mu_1}{z^2}-\frac{\mu_3}{z^3} \\ \frac{\mu_3}{z}+\frac{\mu_2-\mu_3\mu_1+\mu_3\mu_2\mu_1}{z^2}+\frac{\mu_1-\mu_2\mu_1-\mu_3\mu_2}{z^3}-z^{-4} \end{bmatrix}$$

$$\begin{bmatrix} 1+\frac{\mu_2\mu_1-\mu_1+\mu_4\mu_3+\mu_3\mu_2}{z}+\frac{-\mu_2+\mu_4\mu_3\mu_2\mu_1-\mu_4\mu_3\mu_1+\mu_3\mu_1-\mu_3\mu_2\mu_1+\mu_4\mu_2}{z^2}+\frac{-\mu_4\mu_3\mu_2-\mu_3+\mu_4\mu_1-\mu_4\mu_1\mu_2}{z^3}-\frac{\mu_4}{z^4} \\ \frac{\mu_4}{z}+\frac{\mu_4\mu_3\mu_2+\mu_3-\mu_4\mu_1+\mu_4\mu_1\mu_2}{z^2}+\frac{\mu_2-\mu_4\mu_3\mu_2\mu_1+\mu_4\mu_3\mu_1-\mu_3\mu_1+\mu_3\mu_2\mu_1-\mu_4\mu_2}{z^3}+\frac{\mu_1-\mu_2\mu_1-\mu_4\mu_3-\mu_3\mu_2}{z^4}-z^{-5} \end{bmatrix}$$

$$\vdots$$

Figure 3: **Regular tube inverse transfer function growth.** Note the regular increase in the polynomial degrees.

$$\begin{bmatrix} 1 \\ -z^{-3} \end{bmatrix}$$

$$\begin{bmatrix} 1-\frac{\mu_1}{z^3} \\ \frac{\mu_1}{z^2}-z^{-5} \end{bmatrix}$$

$$\begin{bmatrix} 1+\frac{\mu_2\mu_1}{z^2}-\frac{\mu_1}{z^3}-\frac{\mu_2}{z^5} \\ \frac{\mu_2}{z^4}+\frac{\mu_1}{z^6}-\frac{\mu_2\mu_1}{z^7}-z^{-9} \end{bmatrix}$$

$$\begin{bmatrix} 1+\frac{\mu_2\mu_1}{z^2}-\frac{\mu_1}{z^3}+\frac{\mu_3\mu_2}{z^4}-\frac{\mu_2}{z^5}+\frac{\mu_3\mu_1}{z^6}-\frac{\mu_3\mu_2\mu_1}{z^7}-\frac{\mu_3}{z^9} \\ \frac{\mu_3}{z^6}+\frac{\mu_3\mu_2\mu_1}{z^8}-\frac{\mu_3\mu_1}{z^9}+\frac{\mu_2}{z^{10}}-\frac{\mu_3\mu_2}{z^{11}}+\frac{\mu_1}{z^{12}}-\frac{\mu_2\mu_1}{z^{13}}-z^{-15} \end{bmatrix}$$

$$\begin{bmatrix} 1+\frac{\mu_2\mu_1}{z^2}-\frac{\mu_1}{z^3}+\frac{\mu_3\mu_2}{z^4}-\frac{\mu_2}{z^5}+\frac{\mu_4\mu_3+\mu_3\mu_1}{z^6}-\frac{\mu_3\mu_2\mu_1}{z^7}+\frac{\mu_4\mu_3\mu_2\mu_1}{z^8}-\frac{\mu_4\mu_3\mu_1+\mu_3}{z^9}+\frac{\mu_4\mu_2}{z^{10}}-\frac{\mu_4\mu_3\mu_2}{z^{11}}+\frac{\mu_4\mu_1}{z^{12}}-\frac{\mu_4\mu_2\mu_1}{z^{13}}-\frac{\mu_4}{z^{15}} \\ \frac{\mu_4}{z^6}+\frac{\mu_4\mu_2\mu_1}{z^8}-\frac{\mu_4\mu_1}{z^9}+\frac{\mu_4\mu_3\mu_2}{z^{10}}-\frac{\mu_4\mu_2}{z^{11}}+\frac{\mu_4\mu_3\mu_1+\mu_3}{z^{12}}-\frac{\mu_4\mu_3\mu_2\mu_1}{z^{13}}+\frac{\mu_3\mu_2\mu_1}{z^{14}}-\frac{\mu_4\mu_3+\mu_3\mu_1}{z^{15}}+\frac{\mu_2}{z^{16}}-\frac{\mu_3\mu_2}{z^{17}}+\frac{\mu_1}{z^{18}}-\frac{\mu_2\mu_1}{z^{19}}-z^{-21} \end{bmatrix}$$

$$\vdots$$

Figure 4: **DRM tube inverse transfer function growth.** Note the disturbance of the polynomial degrees.

The classical form of the Levinson-Durbin algorithm for AR filter design is described in [RJ93] by :

$$E^{(0)} = r_0 \tag{35}$$

$$\mu_{m+1} = \left[r_{m+1} - \sum_{i=1}^{m} a_i^{(m)} r_{m+1-i} \right] / E^{(m)} \tag{36}$$

$$\left\{ \begin{array}{lcll} a_{m+1}^{(m+1)} & = & \mu_{m+1} & \\ a_i^{(m+1)} & = & a_i^{(m)} + \mu_{m+1} a_{m+1-i}^{(m)} & \quad i = 1, \cdots, m \end{array} \right. \tag{37}$$

$$E^{(m+1)} = \left(1 - \mu_{m+1}^2\right) E^{(m)} \tag{38}$$

with :

- a_m : LPC coefficients
- μ_m : reflection coefficients
- $r_m = \sum_{n=0}^{N-1-m} x(n)x(n+m)$: values of the (estimated) autocorrelation function.

What does $\mu_{m+1} = 0$ brings about the correlation and the LPC coefficients ?

- From equation (37), it simply means that the predictor has not changed between step m and step $m+1$ of the algorithm.
- From equation (38), it means that the energy of the prediction error stays the same.
- From equation (36), setting $\mu_{m+1} = 0$ induces :

$$r_{m+1} = \sum_{i=1}^{m} a_i^{(m)} r_{m+1-i} \tag{39}$$

Adopting a matrix notation, we have :

$$r_{m+1} = \left[a_1^{(m)}, a_2^{(m)}, \cdots a_m^{(m)} \right] \begin{bmatrix} r_m \\ r_{m-1} \\ \vdots \\ r_1 \end{bmatrix} \tag{40}$$

This is a way of constraining the autocorrelation matrix.

Given that :

- the autocorrelation matrix is supposed to be estimated from the original signal
- the LPC coefficients at step (m) are determined and fixed at the previous step

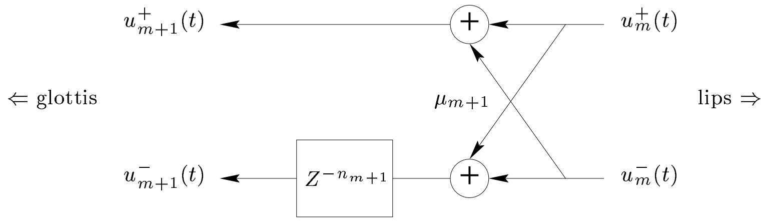
then this constraint amounts to imposing some values in the autocorrelation matrix. We do not just try here to neglect some terms in the matrix. This is therefore equivalent to constraining the original sound signal itself, which we just wish to analyse.

Trying to incorporate the unequal-length tubes constraints in the Levinson recursion leads to the above contradiction. This shows that the original Levinson recursive algorithm cannot be used to solve our problem. Therefore, we are bound to addressing the problem by estimating directly the parameters of the lattice form.

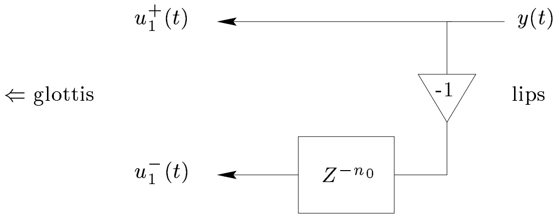
Let us now describe more precisely the lattice form corresponding to the structure of unequal length tubes' transfer functions.

3.3 Lattice forms of the transfer function and its inverse

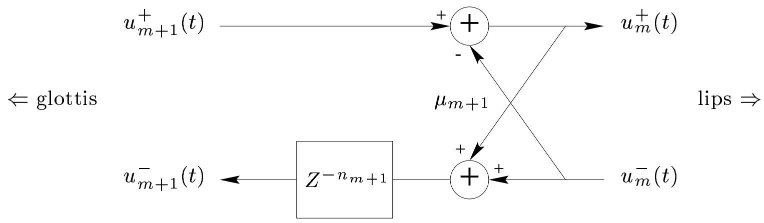
As seen with equation (17), the inverse tube transfer function can be built by connecting serially a number of elementary two-port cells of the form :



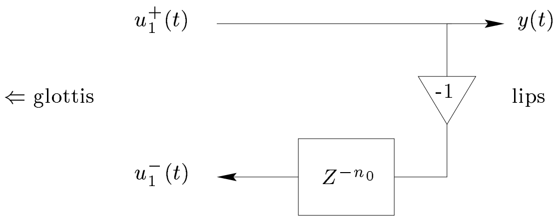
The first cell, at lips end, is of the form :



Alternately, the AR synthesis filter can be represented by a series of cells of the form [OS89] :



with first cell :



Prior to discussing how it is possible to estimate such lattice's parameters, a discussion on filter stability in the unequal lengths tubes case is still necessary.

3.4 Stability of the transfer function in the unequal-lengths tube case

We have shown in section 3.2 that the transfer functions corresponding to synthesis-oriented acoustic tubes had always an Auto-Regressive (AR) form. This form describes Infinite Impulse Response (IIR) filters, which don't have a guaranteed stability. Hence, it is necessary to define the conditions under which the synthesis oriented transfer function $1/D_m^+(z)$ has all its poles inside the unit circle.

It has been demonstrated by Markel and Gray [MG76] and by Wakita [Wak72, Wak73] that equal-length tubes were always stable provided their reflection coefficients $\mu_k,\ k = 1 \cdots m$, had values between -1 and 1 ($\forall k, |\mu_k| < 1$). This condition is equivalent in the physical domain to having positive areas for the tube sections.

It can be seen that the given lattice forms are equivalent to a classical inverse Linear Prediction lattice filter where some reflection coefficients μ_i would be constrained to stay equal to zero. Hence, the constraint introduced by using unequal length means $|\mu_i| = 0$ for some i. Therefore, the stability of the model is still guaranteed if the unconstrained reflection coefficients are between 1 and -1.

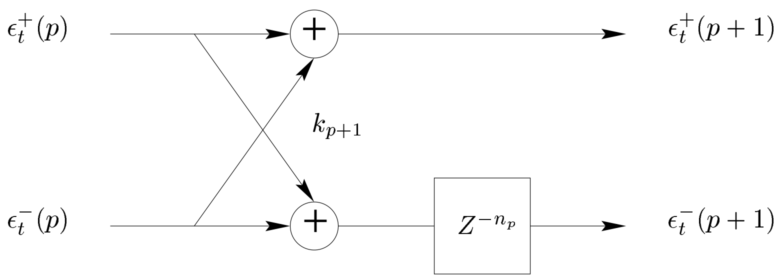
When estimating the transfer function from the acoustic waveform, we will have to verify that this stability condition is theoretically and experimentally respected by the chosen estimation method.

We will now see how it is possible to estimate the parameters of the constrained filters, i.e. estimating the "free" reflection coefficients μ_m from the speech waveform $y(t)$.

4 Application of Burg's method to the DRM inversion

4.1 Inverse DRM lattice filter

As seen in section 3.3, the inverse filtering cell associated to a DRM section corresponds to the following lattice structure:



As we act now in an estimation framework, the notations have been changed to be coherent with the usual notations of the inverse lattice filters estimation theory: estimated reflection coefficients are denoted k_m instead of μ_m, and estimated forward/backward residuals are denoted $\epsilon_t^{+/-}(p)$ at stage p and time t instead of $u_m^{+/-}(t)$.

The forward and backward travelling errors are now characterized by the following relations:

$$\left\{ \begin{array}{lcl} \epsilon_t^+(p+1) & = & \epsilon_t^+(p) \quad + \quad k_{p+1}\, \epsilon_{t-\Delta t_p}^-(p) \\ \epsilon_t^-(p+1) & = & \epsilon_{t-\Delta t_p}^-(p) \quad + \quad k_{p+1}\, \epsilon_t^+(p) \end{array} \right. \tag{41}$$

with: $n_p = \frac{\Delta l_p}{\Delta l_{unit}}$

$\Delta t_p = \frac{\Delta l_p}{c}$

$k_{p+1} = \frac{S_{m+1}-S_m}{S_{m+1}+S_m}$

The complete inverse DRM filter is given in figure 5. It is initialized with the speech waveform y_t by:

$$\left\{ \begin{array}{lcl} \epsilon_t^+(0) & = & y_t \\ \epsilon_t^-(0) & = & -y_t \end{array} \right. \tag{42}$$

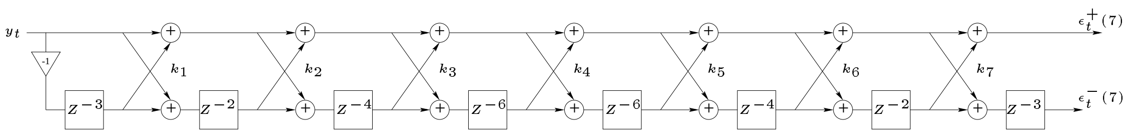


Figure 5: Complete DRM inverse filter.

4.2 Estimation of the inverse filter

To build a correct estimator for the DRM inverse filter, we must define an error criterion and find an analytical solution for its minimization.

4.2.1 Definition of a Least Mean Square error criterion

Let Σ_p denote the sum of all delays from order 1 to order (p):

$$\Sigma_p = \sum_{k=1}^{p} n_k \tag{43}$$

Applying an error criterion similar to the one appearing in Burg's method [Mak77], the mean squared prediction error (MSE) to minimize can be defined as:

$$\xi^2(p+1) = \frac{1}{2}\left\{ \sum_{t=\Sigma_p+1}^{N} \epsilon_t^+(p+1)^2 + \sum_{t=\Sigma_p+1}^{N} \epsilon_t^-(p+1)^2 \right\} \tag{44}$$

i.e., introducing the parameters k_{p+1}:

$$\xi^2(p+1) = \frac{1}{2}\left\{ \sum_{t=\Sigma_p+1}^{N} \left[\epsilon_t^+(p) + k_{p+1}\epsilon_{t-n_p}^-(p)\right]^2 + \sum_{t=\Sigma_p+1}^{N} \left[\epsilon_{t-n_p}^-(p) + k_{p+1}\epsilon_t^+(p)\right]^2 \right\} \tag{45}$$

4.2.2 Minimization of the Mean Squared Error criterion

The correct estimator for the k_i must minimize the mean squared error. Hence:

$$\begin{aligned} \frac{\partial \xi^2(p+1)}{\partial k_{p+1}} &= 0 \\ &= 2\sum_{t=\Sigma_p+1}^{N} \epsilon_t^+(p)\epsilon_{t-n_p}^-(p) + k_{p+1}\left[\sum_{t=\Sigma_p+1}^{N} \left(\epsilon_t^+(p)\right)^2 + \sum_{t=\Sigma_p+1}^{N} \left(\epsilon_{t-n_p}^-(p)\right)^2\right] \end{aligned} \tag{46}$$

$$\Rightarrow \boxed{k_{p+1} = \frac{-2\sum_{t=\Sigma_p+1}^{N} \epsilon_t^+(p)\,\epsilon_{t-n_p}^-(p)}{\sum_{t=\Sigma_p+1}^{N}\left(\epsilon_t^+(p)\right)^2 + \sum_{t=\Sigma_p+1}^{N}\left(\epsilon_{t-n_p}^-(p)\right)^2}} \tag{47}$$

4.3 Stability issues

4.3.1 Verification of the stability condition

To verify that $\forall i$, $|k_i| < 1$, in accordance with the condition defined in section 3.4 and given the form of (47), we must show that $\forall (a,b) \in \mathcal{R}$:

$$\left| \frac{-2\sum a\,b}{\sum a^2 + \sum b^2} \right| < 1 \tag{48}$$

$$\begin{aligned} \Leftrightarrow |-2ab| &< a^2 + b^2 \\ \pm 2ab &< a^2 + b^2 \\ 0 &< a^2 \mp 2ab + b^2 \\ 0 &< (a \pm b)^2 \quad \text{QED} \end{aligned} \tag{49}$$

Given the form of the k_p defined by (47), the condition $\forall i$, $|k_i| < 1$ is always respected by our estimator. Hence filter stability and positive tube sections are guaranteed in theory by our estimation scheme.

4.3.2 Effect of the stability condition

An interesting property of our estimator, related to the filter stability condition, is that the MSE decreases after each cell. Let us express the link between MSE at step $(p+1)$ and MSE at step (p):

$$\begin{aligned}
\xi^2(p+1) &= \frac{1}{2}\left\{\sum_{t=\Sigma_p+1}^{N} \epsilon_t^+(p+1)^2 + \sum_{t=\Sigma_p+1}^{N} \epsilon_t^-(p+1)^2\right\} \\
&= \frac{1}{2}\left\{\sum_{t=\Sigma_p+1}^{N} \left(\epsilon_t^+(p) + k_{p+1}\epsilon_{t-n_p}^-(p)\right)^2 + \sum_{t=\Sigma_p+1}^{N} \left(\epsilon_{t-n_p}^-(p) + k_{p+1}\epsilon_t^+(p)\right)^2\right\} \\
&= \frac{1}{2}\left\{\sum_{t=\Sigma_p+1}^{N} \left(\epsilon_t^+(p)^2 + 2k_{p+1}\epsilon_t^+(p)\epsilon_{t-n_p}^-(p) + k_{p+1}^2\epsilon_{t-n_p}^-(p)^2\right)\right. \\
&\quad \left. + \sum_{t=\Sigma_p+1}^{N} \left(\epsilon_{t-n_p}^-(p)^2 + 2k_{p+1}\epsilon_{t-n_p}^-(p)\epsilon_t^+(p) + k_{p+1}^2\epsilon_t^+(p)^2\right)\right\} \\
&= \xi(p)^2 + \frac{1}{2}\left\{4k_{p+1}\sum_{t=\Sigma_p+1}^{N} \epsilon_t^+(p)\epsilon_{t-n_p}^-(p) + k_{p+1}^2\left(\sum_{t=\Sigma_p+1}^{N} \epsilon_t^+(p)^2 + \sum_{t=\Sigma_p+1}^{N} \epsilon_{t-n_p}^-(p)^2\right)\right\}
\end{aligned} \tag{50}$$

By developing one of the k_{p+1}, using (47), in the second member of the bracketed addition, we obtain:

$$\begin{aligned}
\xi^2(p+1) &= \xi(p)^2 \\
&\quad +2k_{p+1}\sum_{t=\Sigma_p+1}^{N} \epsilon_t^+(p)\epsilon_{t-n_p}^-(p) \\
&\quad +k_{p+1}\frac{-\sum_{t=\Sigma_p+1}^{N}\epsilon_t^+(p)\epsilon_{t-n_p}^-(p)}{\sum_{t=\Sigma_p+1}^{N}\epsilon_t^+(p)^2 + \sum_{t=\Sigma_p+1}^{N}\epsilon_{t-n_p}^-(p)^2}\sum_{t=\Sigma_p+1}^{N}\left(\epsilon_t^+(p)^2 + \epsilon_{t-n_p}^-(p)^2\right) \\
&= \xi(p)^2 - k_{p+1}\sum_{t=\Sigma_p+1}^{N} \epsilon_t^+(p)\epsilon_{t-n_p}^-(p) \\
&= \xi(p)^2 - k_{p+1}\left(k_{p+1}\,\xi(p)\right)
\end{aligned} \tag{51}$$

$$\Rightarrow \boxed{\xi^2(p+1) = (1-k_{p+1}^2)\xi^2(p)} \tag{52}$$

Hence, for each cell added, the MSE is guaranteed to decrease.

5 Experimental results

5.1 Experimental protocol

5.1.1 Goal

The proposed experiments aim at comparing the performances of the DRM inverse filter with performances of two "classical" inverse LPC filters, plus a "constrained" LPC model. The considered models are described in figure 6.

The measure of performance is given by the Mean Squared residual Error (MSE), i.e. the mean of the squared forward residual error after an inverse filtering pass (the lower the MSE, the better the model). The proposed mean is evaluated over complete utterances.

5.1.2 Data set

For our experiments, we have used the sound files available from the University of Wisconsin (UWisc) acoustico-articulatory database [Wes94]. This choice has been made in accordance with other projects led in the framework of articulatory speech recognition at IDIAP. UWisc has been chosen as a reference database to lead articulatory speech recognition experiments.

Sound files in this database have the following characteristics:

Overall characteristics

- *Language*: english
- *Number of speakers*: 48 speakers, 22 males, 26 females
- *Speech volume*:
  - time: approx. 17 minutes per speaker
  - phonemes: approx. 6600 phonemes per speaker with the DARPABET phoneset
- *Nature of the utterances*: prompted text,
  - 40% isolated sentences
  - 33% isolated words and sounds
  - 13% prose (connected speech, long texts)
  - 8% oral motor tasks (jaw wagging, water swallowing,...)
  - 6% isolated numbers

For our experiments, only the isolated words have been used.

Audio

- *Sampling Freq.*: 21739 Hz
- *Format*: NIST SPHERE, "shorten" type compression
- *Quality*: microphone speech, thin band of machine noise around 5435 Hz, some operator speech during silences, some background white noise.

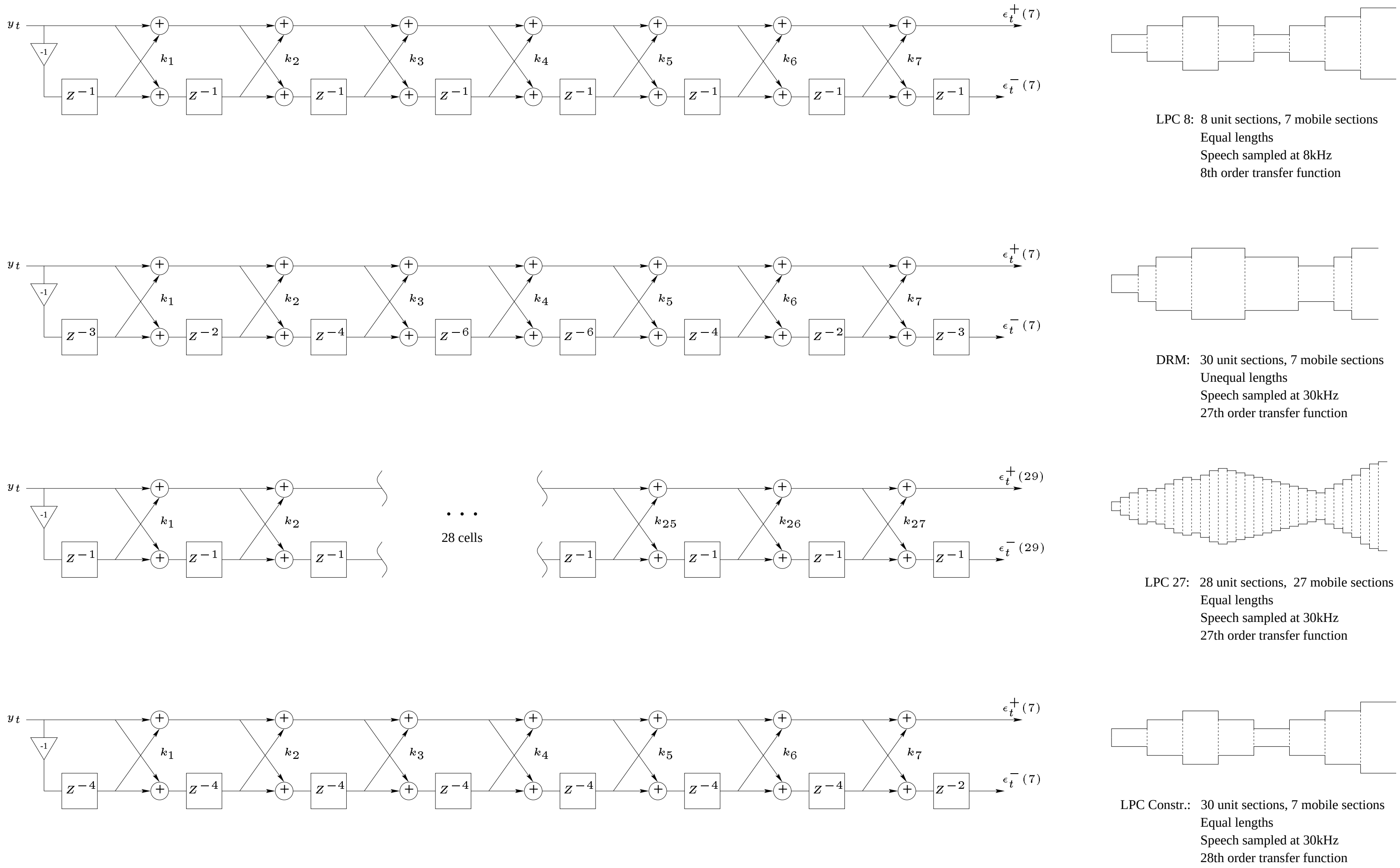


Figure 6: Models compared in our experiments.

5.1.3 Summary of the inverse filtering analysis method

For each of the proposed models, the inverse filtering analysis method is similar and can be decomposed into the following steps:

1. **adapt the sample frequency of speech data** to the constraint imposed by the model, namely:

$$F_s = \frac{c}{2\,\Delta l_{unit}} \tag{53}$$

   where c is the speed of sound ($34000\,cm/s$) and Δl_{unit} is the unit length defined in section 3.1.2. Given the structure of the models, and assuming that a vocal tract is 17 centimeters long on average, the sampling frequencies to use will be 8kHz in the case of LPC8 and 30kHz in the cases of the DRM, LPC27 and LPC Constr.

   The frequency adaptation can be performed from the unaltered original 21kHz sampling rate, or after a low-pass filtering with cutoff at 4kHz to simulate telephone speech. The polyphase filtering method [BBC76, Ell87] applied to resampling has been chosen for its computational efficiency alied to good interpolation quality (little or no aliasing is introduced by this method). The low-pass/4kHz cutoff filtering has been applied directly in the framework of the polyphase filtering scheme when needed.

2. **pre-emphasize the obtained speech wave** by a simple differentiation.

   This step is performed to compensate for the effects introduced by the glottal waveform shape and the radiation effect at the lips. These effects are not otherwise modelled by our inverse filters.

3. **inverse-filter speech**. Adapt filter every 10ms by computing reflection coefficients k_i, using expression (47) with 25ms observation windows.

4. **compute mean squared error** in the different cases studied.

5.2 Analysis of the results

5.2.1 Influence of the model

The development of the DRM analysis method leads to the estimation of a constrained 27th order FIR inverse filter, characterized by 7 parameters instead of 27 coefficients.

Results depicted in figure 7 show that with the same number of models' parameters, the DRM-derived analysis results in a lower prediction error than analysis derived from an equal-length tube model: the DRM performs better than the classical 8th order LPC corresponding to a tube with 7 reflection coefficients.

As can be logically expected, a classical (equal-length tube) LPC of an order equivalent to that of the DRM leads to a lower prediction error because it does not incorporate particular constraints: LPC analysis of order 27 on 30kHz speech performs better than the DRM.

It can also be seen that the particular repartition of the section lengths in the DRM plays a role in the reduction of the prediction error: the LPC Constr. tube, with 7 mobile equal-length sections equivalent to a 28th order LPC on 30kHz speech (see figure 6), performs worse than the DRM. Hence, it can be inferred that the improvement brought by the DRM does not come only from a better estimation of the k_i, which would arise from the higher quantity of data present in the 25ms observation window when analyzing speech at 30kHz instead of 8kHz. The repartition of the sections lengths itself appears to play a significant role in the reduction of the prediction error.

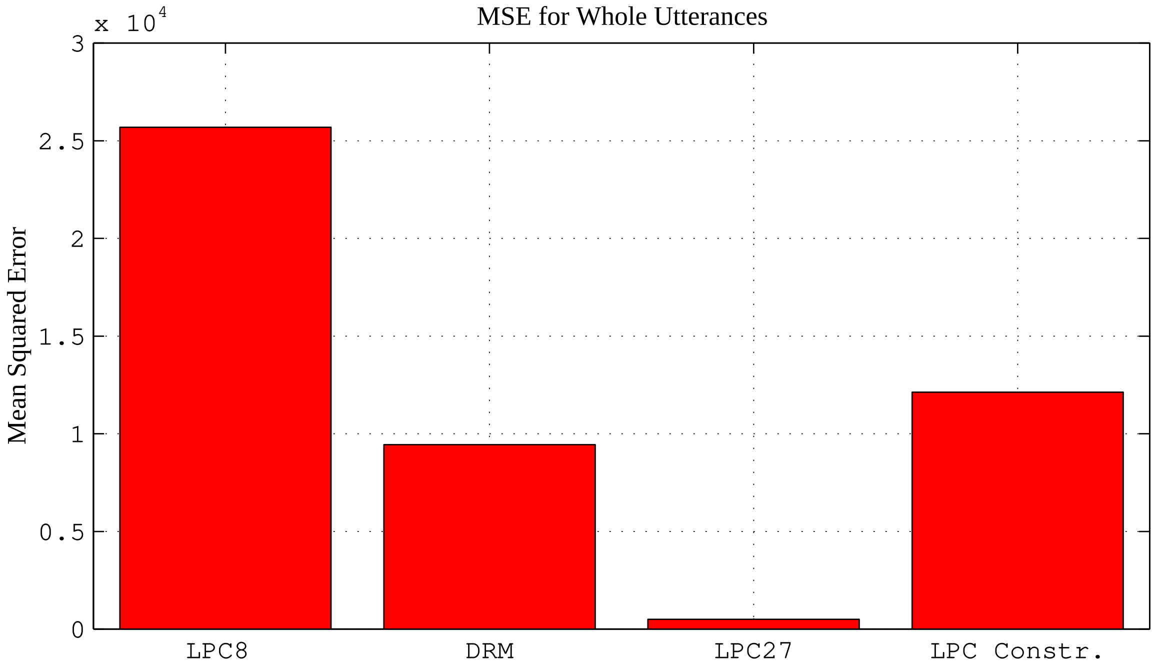


Figure 7: Comparison of Mean Squared Errors for the different tube models. Upsampling or downsampling is performed from the original 21kHz sample frequency.

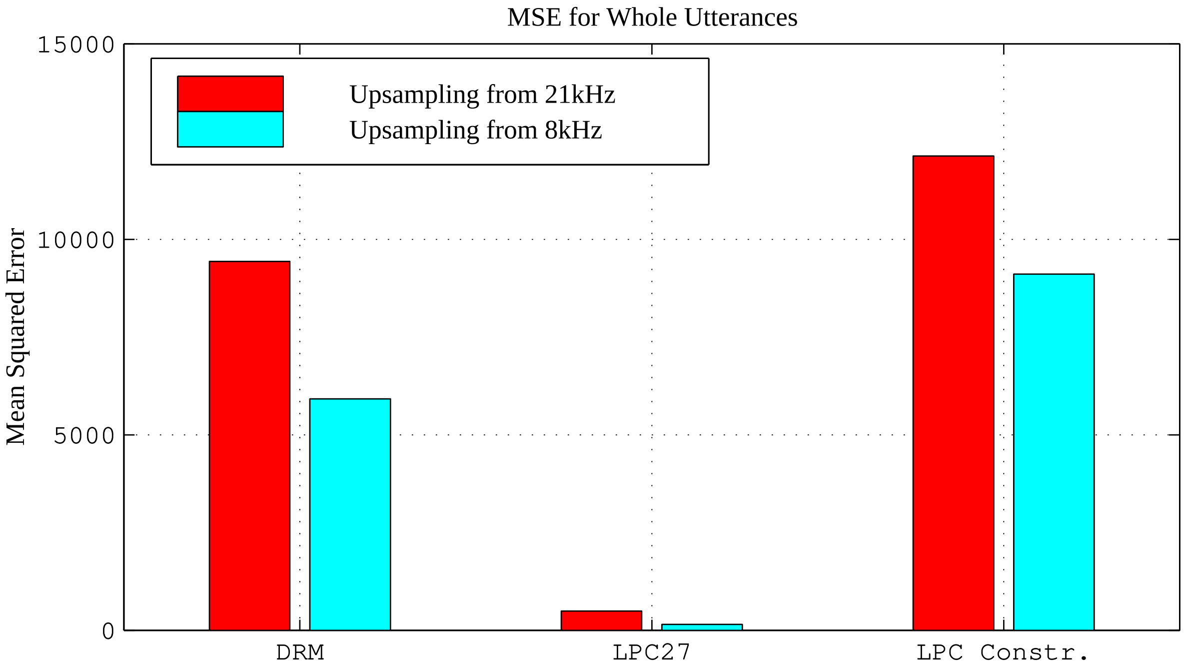


Figure 8: Comparison of Mean Squared Errors for different tube models, upsampling from two different speech qualities (21kHz and 8kHz sampling rates).

5.2.2 Influence of band-limiting the speech waveform

Figure 8 shows that band-limiting the speech waveform to 0-4kHz helps further reducing the prediction error. This fact is to be related to the natural tendency of LPC analysis to over-estimate the amplitude of the higher spectral peaks. This property is also the expression of one of the fluid dynamics assumptions evoked in section 3.1.1 : tube quantization introduces acceptable error if lengths of the sections are kept short compared to a wavelength at the highest frequency of interest ([Fla72], p.25), i.e. $L_{max} \ll \frac{c}{F_{max}}$ where c is speed of sound. As a matter of fact, when band-limiting to 0∼4kHz, the longest wavelength is kept well above the longest tube section. In opposite, when using the whole 0∼10kHz bandwidth, the longest wavelength comes closer to the longest tube section, thus allowing for a higher modeling error.

Highest frequency	Corresponding Wavelength ($c = 34000\,cm/s$)
21kHz /2	$\approx$ 3 cm
8kHz /2	8.5 cm

Model	L_{max} (assuming a vocal tract of 17cm)
DRM	3.4 cm
LPC30	$\approx$ 0.6 cm
LPC Constr.	$\approx$ 2.3 cm

6 Conclusions

The present work was concerned with the construction of a speech analysis method incorporating constraints inherited from the DRM speech production model. Those constraints have been expressed in the context of lattice filters' parameters estimation. Experiments have shown that with an equal number of estimated parameters, the DRM-derived method was bringing an improvement over classical Auto-Regressive analysis in terms of lower modeling error.

Future work of interest concerns the application of this analysis method in the areas of speech coding and speech recognition. A validation for speech coding would require a study of the sensitivity of the DRM-related reflection coefficients and residual signal to the effects of quantization. An application to speech recognition would suppose that a method to compute cepstral coefficients from the DRM-related reflection coefficients is laid out, and that the resulting coefficients are tested in a speech recognition system. Future work from our part will be directed along the second axis.

References

[BBC76] M. Bellanger, G. Bonnerot, and M. Coudreuse. Digital filtering by polyphase network: application to sample rate alteration and filter banks. *IEEE transactions on Acoustics, Speech and Signal Processing*, ASSP-24(2), April 1976.

[Bon83] L.J. Bonder. The n-tube formula and some of its consequences. *Acustica*, 52:216–226, 1983.

[Che95] S. Chennoukh. *Modélisation du conduit vocal en régions distinctives. Synthèse d'ensembles Voyelle-Voyelle et Voyelle-Consonne-Voyelle.* PhD thesis, ENST, mars 1995.

[Ell87] Douglas F. Elliott, editor. *Handbook of Digital Signal Processing.* Academic Press, 1987.

[Fan73] G. Fant. *Speech Sounds and Features.* MIT Press, Cambridge, 1973.

[Fla72] J.L. Flanagan. *Speech analysis, synthesis and perception.* Springer Verlag, 1972.

[FP74] G. Fant and S. Pauli. Spatial characteristics of vocal tract resonance modes. Speech Communication Seminar, 1974.

[Mak77] J. Makhoul. Stable and efficient lattice methods for linear prediction. *IEEE trans. on Acoustics, Speech and Signal Processing*, ASSP-25(5):423–428, October 1977.

[MCG88] M. Mrayati, R. Carré, and B. Guérin. Distinctive regions and modes: a new theory of speech production. *Speech Communication*, (7):257–286, 1988.

[MG76] J.D. Markel and A.H. Gray. *Linear prediction of speech.* Springer-Verlag, 1976.

[MI68] P.M. Morse and K.U. Ingard. *Theoretical acoustics.* Mc Graw-Hill, 1968.

[OS89] A.V. Oppenheim and R.W. Schafer. *Discrete time signal processing.* Prentice Hall, 1989.

[RJ93] L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[Wak72] H. Wakita. Estimation of vocal tract shape by optimal inverse filtering and acoustic/articulatory conversion methods. Technical Report 9, Speech Communication Research Lab, Santa Barbara, Cal., 1972.

[Wak73] H. Wakita. Direct estimation of the vocal-tract shape by inverse filtering of acoustic speech waveforms. *IEEE Transactions on Audio and Electroacoustics*, AU-21:417–427, October 1973.

[Wes94] J.H. Westbury. *X-ray microbeam speech production database user's handbook.* Waisman Center, University of Wisconsin, 1.0 edition, June 1994.