INVERSE LATTICE FILTERING OF SPEECH WITH ADAPTED NON-UNIFORM DELAYS

Sacha KRSTULOVIĆ [1] *Frédéric BIMBOT* [2]

[1] IDIAP C.P. 592 - CH-1920 Martigny - Switzerland - sacha@idiap.ch
[2] IRISA - Campus Beaulieu, 35042 Rennes, France - bimbot@irisa.fr

ABSTRACT

A particular form of constraint is incorporated to Linear Prediction lattice filter models in the form of unequal-length delays. This constraint amounts to reducing the number of intrinsic degrees of freedom defined by the reflection coefficients without modifying the LPC order of the corresponding transfer function. It can be optimized by a simple exhaustive search scheme. Preliminary results show that the prediction error is slightly decreased with respect to a conventional predictor using the same number of reflection coefficients.

1. INTRODUCTION

It is well known [MG76, Wak73] that the process of AR filtering is equivalent, under certain hypothesis, to the acoustic filtering in discrete lossless tubes. While this is traditionally established for tubes with discrete sections of even unitary lengths, an extension concerning the case of tubes with a non-uniform repartition of lengths has been formulated recently [Krs98]. Non-uniform lengths, or equivalently non-uniform delays in a lattice filter, form a particular kind of production constraint.

This article presents a first series of experiments aiming at observing the modifications of inverse filtering performance induced by the incorporation of this constraint into a classical Auto-Regressive modeling framework. After having reviewed the theory of non-uniform lattice filters and their estimation, we will describe a method to optimize the tube configuration in order to minimize the inverse filtering residual error. Experimental results pertaining to the application of the method to various signals will then be given and discussed.

2. REDUCTION OF DEGREES OF FREEDOM IN AN AR MODEL

2.1. AR models and acoustic tubes

The transfer function $A(z) = \frac{1}{D^+_M(z)}$ of a lossless tube discretized in M unequally lengthy individual sections can be computed in a digital processing framework by application of the following matrix recursion [Krs98]:

$$\begin{bmatrix} D^+_{m+1}(z) \\ D^-_{m+1}(z) \end{bmatrix} = \begin{bmatrix} 1 & k_{m+1} \\ k_{m+1} z^{-n_{m+1}} & z^{-n_{m+1}} \end{bmatrix} \begin{bmatrix} D^+_m(z) \\ D^-_m(z) \end{bmatrix}$$

$$\begin{bmatrix} D^+_0(z) \\ D^-_0(z) \end{bmatrix} = \begin{bmatrix} 1 \\ -z^{-n_0} \end{bmatrix} \qquad (1)$$

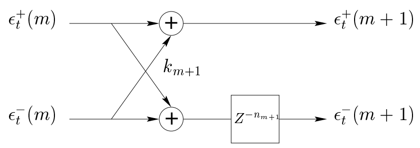
where:

- $D^+_m(z)$ is the inverse transfer function for the forward traveling sound wave after connection of the m^{th} tube section ($D^-_m(z)$ being the transfer function for the backward traveling wave)
- k_{m+1} is the reflection coefficient between section (m) and section $(m+1)$, having areas S_m and S_{m+1} respectively. It is defined as $k_{m+1} = \frac{S_{m+1} - S_m}{S_{m+1} + S_m}$.
- n_m is the delay order of section m, with variable z defined as $z = \exp^{j\omega 2\Delta_{unit} t}$. $\Delta_{unit}t$ is the time necessary for sound to travel along one unit of length of a tube section. The unit length is defined as the greatest common divisor of the sections' lengths. The definition of the corresponding $\Delta_{unit}t$ time constant is where the connection between physical measures in a tube and dimensionless signal processing operates [Krs98].

Equation (1) corresponds to a lattice filtering structure where each matrix

$$\begin{bmatrix} 1 & k_{m+1} \\ k_{m+1} z^{-n_{m+1}} & z^{-n_{m+1}} \end{bmatrix} \qquad (2)$$

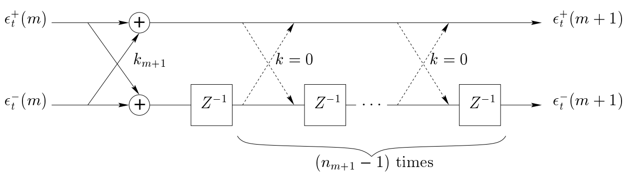
corresponds to an inverse filtering cell of the form:



(see figure 1 for a complete filter). When the delays n_{m+1} are non-unit, these matrices can be expanded as:

$$\underbrace{\begin{bmatrix} 1 & 0 \\ 0 & z^{-1} \end{bmatrix} \cdots \begin{bmatrix} 1 & 0 \\ 0 & z^{-1} \end{bmatrix}}_{(n_{m+1}-1) \text{ times}} \begin{bmatrix} 1 & k_{m+1} \\ k_{m+1} z^{-1} & z^{-1} \end{bmatrix} \qquad (3)$$

i.e., in the lattice form:



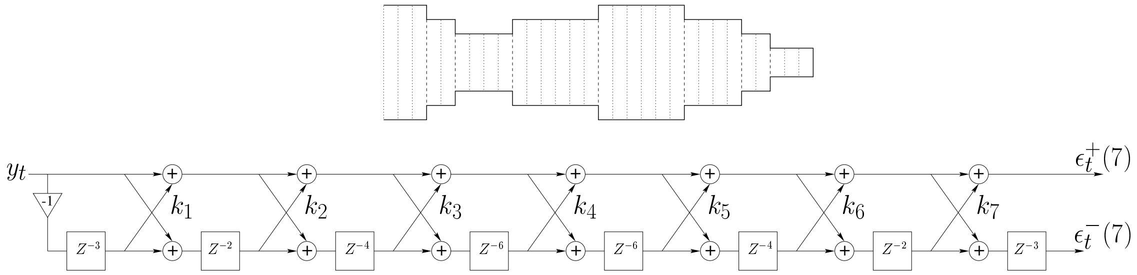


Figure 1: Inverse lattice filter accounting for a particular tube configuration, with 30 units sections / 7 degrees of freedom. In this example, 22 reflection coefficients are fixed to zero, and 7 are "free".

Hence, the unequal length delays topology is equivalent to constraining some reflection coefficients to zero in a filter with a given number of unit-length delays.

It can be verified that the function $D_M^+(z)$ resulting from the application of recursion (1) is always a polynom in z^{-1}. Hence, $A(z) = \frac{1}{D_M^+(z)}$ is an Auto-Regressive transfer function. In the classical unit-delays case, the order of the transfer function is equal to the number of reflection coefficients used to build it. Conversely, in the constrained case, the order of $A(z)$ is the sum of the non-unit delays n_m. This order is equal to the number of unit sections in the corresponding tube model, and it is greater than the number of reflection coefficients used to parameterize the function since the number of free reflection coefficients is only equal to the number of unequal-length sections in the tube. Hence, using non-uniform delays amounts to imposing an intrinsic number of degrees of freedom (DoFs) in a LPC model with a given order. Alternately, it allows to predict a signal sample from an increased portion of its past while keeping the number of model parameters fixed.

The stability of the constrained filter is preserved since forcing some reflection coefficients k_i to zero respects the general stability condition for a lattice filter [MG76], namely $|k_i| < 1 \ \forall i$ (every k_i should have a value between -1 and 1).

2.2. Modeling with non-uniform filters

The parameters k_i of the constrained filters can be estimated by expressing the constraints into Burg's estimator [Mak77]. Denoting by Σ_p the sum of all the delays from order 1 to order (p), Burg's mean squared error criterion can be expressed as:

$$\xi^2(p+1) = \frac{1}{2}\left\{ \sum_{t=\Sigma_p+1}^{N} \epsilon_t^+(p+1)^2 + \sum_{t=\Sigma_p+1}^{N} \epsilon_t^-(p+1)^2 \right\} \quad (4)$$

where $\epsilon_t^+(p)$ (resp. $\epsilon_t^-(p)$) is the inverse filtering residual at the output of the forward (resp. backward) branch of the lattice filter. Minimizing this error criterion through differentiating and equating to zero gives:

$$k_{p+1} = \frac{-2\sum_{t=\Sigma_p+1}^{N} \epsilon_t^+(p)\epsilon_{t-n_p}^-(p)}{\sum_{t=\Sigma_p+1}^{N} \left(\epsilon_t^+(p)\right)^2 + \sum_{t=\Sigma_p+1}^{N} \left(\epsilon_{t-n_p}^-(p)\right)^2} \quad (5)$$

It can be easily verified that this estimator always generates values that lie between -1 and 1, and hence always produces stable filters.

2.3. Detail of the analysis method

Our modeling method is similar to the classical frame-based analysis, but using the modified estimator:

1. *pre-emphasize the input speech signal.* This classical step is performed to compensate for the effects introduced by the glottal waveform shape and the radiation effect at the lips, in the hope that the estimated parameters will better capture some vocal tract properties.
2. *estimate the filter coefficients* every 10ms by computing reflection coefficients k_i, using expression (5) with 25ms observation windows. The extracted reflection coefficients can be further transformed into log area ratios or area functions.
3. *inverse-filter the signal and compute the Mean Squared prediction Error (MSE)* pertaining to the input data and the tested model:

$$MSE = \frac{1}{2N}\sum_{i=1}^{N}\left(\left(\epsilon_i^+(M)\right)^2 + \left(\epsilon_i^-(M)\right)^2\right) \quad (6)$$

2.4. Optimization

Various constraints lead to different filtering performances in terms of a higher or lower MSE for a test signal. Hence, it is interesting to find the best performing topology given a number of degrees of freedom to be distributed over a given total LPC order.

To search for this best configuration, all the lattice filter topologies respecting a given number of DoFs are generated and systematically used to inverse-filter a test sentence. The one bringing the least MSE is regarded as the best topology.

The various configurations are identified by strings of the form, e.g., [5/22:3x3,8,5.]. This example reads: "5 unequal-length sections distributed on a 22 unit sections model, with three 3^{rd} order delays, one 8^{th} order delay and one 5^{th} order delay".

#DoFs	# conf.	Male speaker 32 kHz	
1 DoF	31	2/32:2,30.	85104
2 DoF	465	3/32:1,1,30.	59386
3 DoF	4'495	4/32:1,1,2,28.	39750
4 DoF	31'465	5/32:1,1,2,2,26.	31562
5 DoF	169'911	6/32:4x1,2,26.	28859
...	> 700'000		
26 DoF	169'911	27/32:18x1,3,6x1,2,3.	20631
27 DoF	31'465	28/32:19x1,2,6x1,2,3.	20531
28 DoF	4'495	29/32:27x1,2,3.	20446
29 DoF	465	30/32:29x1,3.	20358
30 DoF	31	31/32:29x1,2,1.	20292
31 DoF	1	32/32:32x1.	20273
Constraints including minimum length :			
6 DoF minlen = 2	134'596	7/32:5x2,18,4.	42764
7 DoF minlen = 2	245'157	8/32:6x2,16,4.	41281

#DoFs	# conf.	Speaker jw11 (m)		Speaker jw16 (f)	
1 DoF	21	2/22:1,21.	2363	2/22:1,21.	24092
2 DoF	210	3/22:1,1,20.	1283	3/22:1,1,20.	10953
3 DoF	1'330	4/22:1,1,5,15.	1115	4/22:1,1,2,18.	8425
4 DoF	5'985	5/22:1,1,4,4,12.	1002	5/22:1,1,2,2,16.	7471
5 DoF	20'349	6/22:1,1,2,2,1,15.	923	6/22:4x1,2,16.	6689
6 DoF	54'264	7/22:1,1,2,2,1,4,11.	821	7/22:4x1,2,2,14.	6313
7 DoF	116'280	8/22:1,1,2,2,1,3,1,11.	757	8/22:4x1,3x2,12.	5927
8 DoF	203'490	9/22:4x1,2,1,3,1,11.	715	9/22:6x1,2,2,12.	5574
...					
14 DoF	116'280	15/22:11x1,3,2,2,4.	554		
15 DoF	54'264	16/22:11x1,3,2,1,1,4.	542	16/22:10x1,3x2,1,3,2.	4556
16 DoF	20'349	17/22:11x1,2,1,2,1,1,4.	530	17/22:10x1,2,2,3x1,3,2.	4460
17 DoF	5'985	18/22:14x1,2,1,1,4.	520	18/22:12x1,2,3x1,3,2.	4369
18 DoF	1'330	19/22:14x1,2,1,1,3,1.	510	19/22:17x1,3,2.	4301
19 DoF	210	20/22:14x1,2,1,1,2,1,1.	501	20/22:17x1,3,1,1.	4251
20 DoF	21	21/22:18x1,2,1,1.	494	21/22:18x1,2,1,1.	4203
21 DoF	1	22/22:22x1.	487	22/22:22x1.	4171

Table 1: Best configurations for various constraints applied to 32^{nd} and 22^{nd} order models (with corresponding MSE).

Further constraints, such as a minimum delay order, can be imposed to the optimization scheme to make the number of tested filters more tractable. For instance, in the case of the distribution of 7 degrees of freedom (7 non-null k_i's delimiting 8 sections) on a 31^{st} order LPC process (a 32 unit sections tube), 2'629'575 filters have to be tested. Imposing the minimum delay to be no shorter than 2 units reduces this number to 245'157 filters. Careful design of the test program allows to achieve computation in a tractable time (from about ten hours to a few days).

The results presented here have been computed from three speech signals. One is a portion of a French test sentence ("*dans cette crémerie*") spoken by a male speaker, recorded in an anechoic chamber and sampled at 32kHz. The two others are 3 unconnected words ("*dormitory school has*"), taken from the English word lists of the University of Wisconsin database [Wes94] for a male speaker (jw11) and a female speaker (jw16), recorded in laboratory conditions and sampled at 21.739kHz.

3. EXPERIMENTAL RESULTS

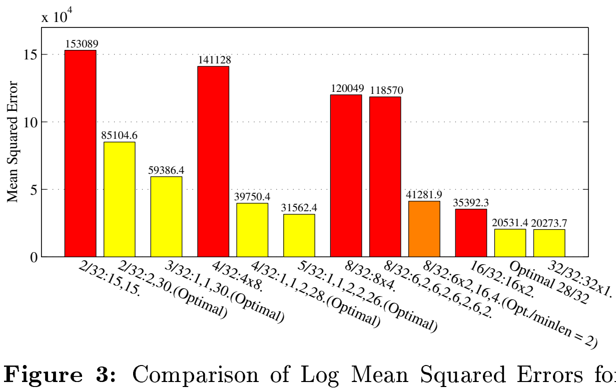


Figure 3: Comparison of Log Mean Squared Errors for various tube models (male speaker sampled at 32kHz).

3.1. Importance of the optimization

Figure 3 shows the MSE after the application of variously constrained filters on the test sentence of the male/32kHz speaker. It shows that the optimization of the constraint brings significant MSE reductions for a given number of degrees of freedom. It also shows, as could be expected, that the best error (light bars) decreases monotonically with the increase of the number of preserved degrees of freedom.

3.2. Optimal configurations

Table 1 gives the optimal configuration for various constraints and for the three tested speech signals. The number of configurations to be tested for each constraint is also indicated. The prediction performance with a reduced number of degrees of freedom remains comparable with the error of unconstrained predictors comprising the same number of DoFs.

3.3. MSE decrease

The MSE comparison is shown in figure 2. It indicates that optimal constrained filters produce a lower MSE than unconstrained filters with an equal number of DoFs. The observed MSE reductions typically range from a few percent to about 30%.

Two different phenomena play a role in these MSE variations. In the case of the unconstrained predictors, the decrease in modeling accuracy is induced by the reduction of the prediction time span (equivalent to the reduction of the LPC order). In the case of the constrained predictors, the decrease in MSE results from the reduction of the number of DoFs, while the prediction time span stays the same. Hence, a `[8/32]` predictor still uses 32 speech samples for its prediction, whereas a `[8/8]` predictor uses only 8 samples. Given that the two curves necessarily join at their lower end (`[32/32]` or `[22/22]`), it appears that the loss induced by the time span reduction is slightly greater than the loss induced by the DoFs reduction. Nevertheless, it must be noted that the error decrease is observed with the very data which has been used for filter optimization : comparison should also be performed on a test set not seen during the optimization stage.

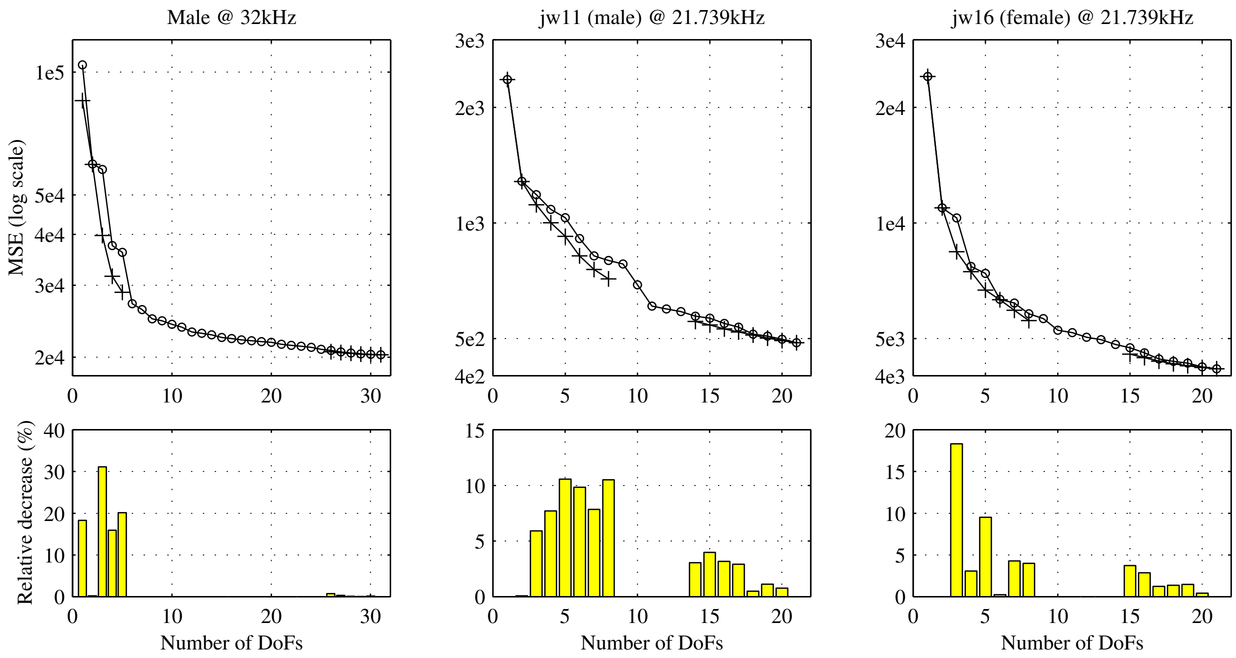


Figure 2: *Top curve*: evolution of the Mean Squared Error for constrained tubes (+) and unconstrained tubes (o) in function of the number of degrees of freedom. *Bottom curve*: relative MSE decrease.

3.4. Dependence upon the signal

Non-speech test signals, scaled to various means and standard deviation values, have yielded different length repartitions (table 2). This confirms that the optimally constrained filters are somehow specialized to the nature of the signals used during optimization.

Signal	Best configuration
pulse train, 100Hz	8/32:17,3,2,2,2,2,2,2.
sine wave, 100Hz	8/32:2,2,15,2,2,5,2,2.
white noise	8/32:2,2,2,2,2,2,2,18.

Table 2: Best configurations for non-speech test signals with an 7 DoFs / 32^{nd} order LPC / minimum length = 2 constraint.

4. CONCLUSION

This article has presented a constrained parametric signal analysis scheme derived from Linear Prediction put in parallel with non-uniform acoustic tube models. This scheme allows to reduce the number of required LPC modeling parameters while keeping the related increase of prediction error to its lowest level. Equivalently, it allows to predict a signal sample from a longer portion of its past with fewer parameters than the usual Auto-Regressive models. This constitutes a preliminary study of a scheme which has a wide variety of potential applications, in particular speech recognition or speech coding.

ACKNOWLEDGMENTS

This work is supported by the Swiss National Science Foundation, grant nr. 20-55.634.98 for the ARTIST II project.

5. REFERENCES

[Krs98] S. Krstulović. Acoustico-articulatory inversion of the DRM model through inverse filtering. IDIAP-RR 16, IDIAP, 1998. [1].

[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.

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

[Wak73] H. Wakita. Direct estimation of the vocal-tract shape by inverse filtering of acoustic speech waveforms. *IEEE Transactions on Audio and Electroacoustics*, pages 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.

[1] Available on
`ftp://ftp.idiap.ch/pub/reports/1998/rr98-16.ps.gz`