RELATING LPC MODELING TO A FACTOR-BASED ARTICULATORY MODEL

Sacha Krstulović

IDIAP C.P. 592 - CH-1920 Martigny - Switzerland - `sacha@idiap.ch`

ABSTRACT

This paper proposes a method for recovering the articulatory parameters of a factor-based vocal tract shape model from the speech waveform. This is realized by analytically relating the shape model to a Linear Prediction lattice filter. Results pertaining to human vowels are presented. They show a good agreement with phonetic characteristics in a real-time computational framework.

1. INTRODUCTION

Most of the existing articulatory models use an area function (describing a discrete tube shape) as the interface between the articulatory level and the acoustics of speech. Alternately, it is well known [MG76, Wak79] that the process of Linear Prediction (LPC) is equivalent, under certain hypotheses, to acoustic filtering in discrete tubes. This equivalence has already been exploited to recover area functions from acoustics alone [Wak79]. In the present article, this method is extended by two further steps, namely:

- projecting the area function into the space of sagittal cuts through the analytic inversion of the $\alpha\beta$ transform [Sun69];
- smoothing the obtained shapes through least-squares decomposition into a basis of shape factors drawn from the statistical analysis of human X-ray data [Mae79].

This amounts to interfacing a linear vocal tract profile model with well known LPC methods in order to achieve real-time acoustico-articulatory inversion. After having reviewed the various components of the corresponding processing chain, we will report some inversion results obtained with both synthetic and human vowels.

2. DESCRIPTION OF THE SYSTEM'S COMPONENTS

The system decomposes into the blocks depicted in figure 1. Each block will be described below.

2.1. Relation between sound and acoustic parameters

Among the available formulations of Linear Prediction modeling (also known as All-Pole modeling), Inverse Lattice Filtering (ILF) [Mak77] occupies a place of choice since: *1)* it provides stable filters, *2)* it does not require a windowing of the input signal, and *3)* the parameters it provides, called reflection coefficients, offer good quantization properties. This method is widely used for speech coding. It decomposes into the following steps:

1. the application of pre-emphasis to digitized speech;
2. the estimation of the reflection coefficients every 10 milliseconds by application of an adequate estimator (Itakura-Saito and Burg being the most widely used [Mak77]), using observation windows of length 25 milliseconds;
3. the inverse filtering of speech, delivering a residual error signal.

For the present system, a 7^{th} order filter has been used with speech sampled at 8kHz. Its 7 reflection coefficients have been estimated with the Itakura-Saito estimator, which minimizes a likelihood distortion between the modeled spectrum and a theoretic optimal All-Pole spectrum.

The reconstruction or synthesis of a speech signal corresponds to exciting a lattice filter with the residual error obtained after inversion, with quantized error sequences (such as in Code Excited Linear Prediction, CELP) or with a white noise, a pulse train or a more elaborate synthetic glottal-like excitation (e.g. Rosenberg's glottal wave). The filter's parameters are updated every 10 milliseconds.

In the synthesis direction, more elaborate acoustic models also exist (e.g. electrical analogies comprising energy loss models). They could be plugged in place of the employed lossless All-Pole model, but they should admit an inversion method to preserve the integration of inversion and synthesis in a unified processing framework. LPC models may be less accurate, but they readily allow a wide range of efficient inversion algorithms.

2.2. Relation between acoustic parameters and area function

Several authors [MG76, Wak79] have shown that the process of All-Pole filtering is analogous to acoustic filtering in discrete lossless acoustic tubes provided that:

1. sound waves are considered to be plane fluid waves,
2. the lengths of the individual tube sections are kept short compared to the wavelength at the highest frequency of interest (this introduces a spectral boundary),
3. the sampling rate of the speech signal is $F_s = \frac{c}{2\Delta l_{unit}}$, where Δl_{unit} is the length of a tube section,
4. no losses are accounted for.

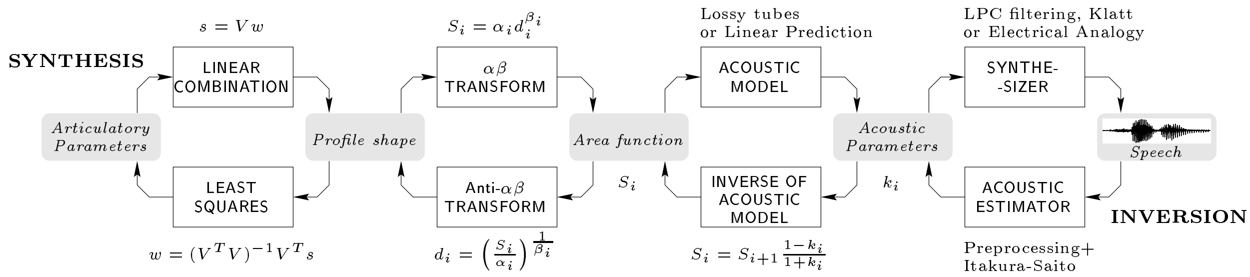

Figure 1: The implemented speech processing chain.

If the speech signal is pre-emphasized to compensate for the spectral characteristics of the glottal excitation and for the radiation impedance at the lips, estimates of vocal tract area functions can be recovered from the speech waveform by using ILF and the following relation :

$$k_i = \frac{S_{i+1} - S_i}{S_{i+1} + S_i} \quad \Leftrightarrow \quad S_i = S_{i+1}\frac{1-k_i}{1+k_i} \qquad (1)$$

where k_i denotes reflection coefficients and S_i denotes the areas of the corresponding discrete lossless tube, numbered in ascending order from lips to glottis. If the lips section is not available, this recursion can be applied by considering the glottis section to be fixed to $1.5cm^2$.

Considering that the area function (or vocal tract) should be 17.5 centimeters long, condition 3 imposes to use 8 sections, or equivalently 7 mobile interfaces, for speech sampled at 8kHz. This imposes the 7^{th} order filter employed in the LPC analysis of section 2.1.

2.3. Connecting areas and profiles

Since the human vocal tract does not have circular sections, the relation between area functions and vocal tract profiles is described by the $\alpha\beta$ transform [Sun69] :

$$S_i = \alpha_i d_i^{\beta_i} \quad \Leftrightarrow \quad d_i = \left(\frac{S_i}{\alpha_i}\right)^{1/\beta_i} \qquad (2)$$

where S_i is the area of a section, d_i is the diameter measured from the profile outline, and (α_i, β_i) are section-dependent parameters. As shown above, this relation admits an exact, one-to-one reciprocal.

Various definitions exist for the diameters d_i. While the works related to Maeda's model usually employ a pseudo-diameter derived from lateral areas, our choice has been to stick to the original $\alpha\beta$ rationale by measuring the diameters along the lines of a semipolar grid. However, area to profile transformations is still an active field of research : numerous other transformations exist [LS96]. Other models can readily replace the original $\alpha\beta$ relation in the processing chain of figure 1 if they prove to be more accurate and still invertible.

Before transformation, the area function may be resampled to meet further processing requirements. In our case, the 8 sections corresponding to the 7^{th} order LPC model have been redistributed over 30 sections to match the dimensions of the profile shape model described in the next section.

2.4. Linear profile shape model

Original model - Maeda's model [Mae79] represents vocal tract profile shapes from 32 measurements made in 3 distinct zones of the vocal tract, using a semipolar grid (fig. 2) :

- in the lips zone, lip aperture ($\mathrm{LIP_{ap}}$), lip protrusion ($\mathrm{LIP_{pr}}$) and lip width ($\mathrm{LIP_{wd}}$) are measured;
- in the tongue region, 25 tongue shape measures are plotted along the semipolar grid lines ($\mathrm{TNG_1}$ to $\mathrm{TNG_{25}}$);
- in the larynx zone, two points delimiting the lower larynx edge are plotted (LRX_{x_1}, etc.).

In addition, a fixed back wall outline is measured in the semipolar grid. It delimits diameters in the lips and tongue regions.

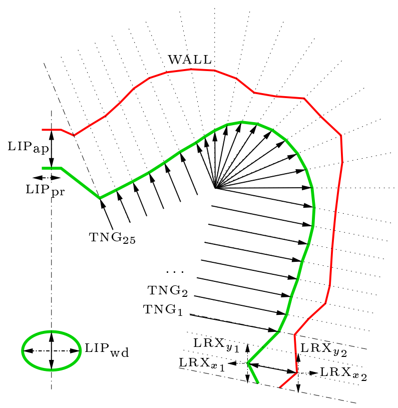

Figure 2: Components of the linear profile shape model.

Each of the mobile features has been related to a set of 5 control parameters by orthogonal factor analysis (a form of driven linear regression) performed on lateral X-ray pictures of a template speaker [Mae79]. The linear factors have been determined so that the control parameters have an articulatory interpretation :

- jw represents the influence of the jaw on all the features
- tp, ts and tt control the tongue position, tongue shape and tongue tip position respectively
- lh and lp control the lip height and the lip protrusion
- lx controls the larynx height.

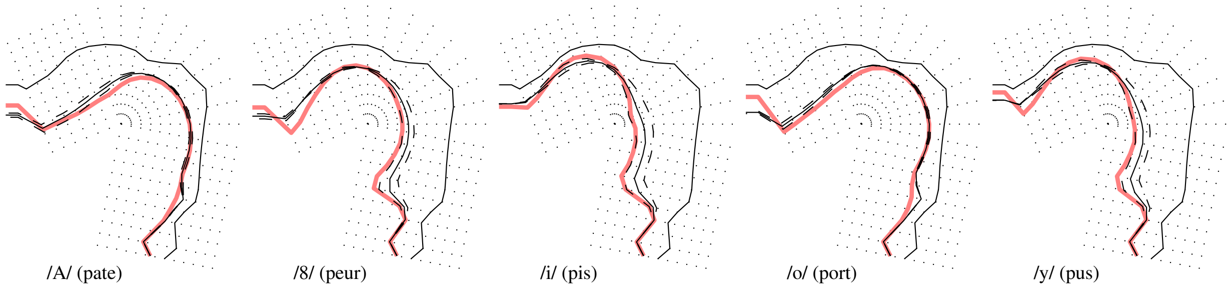

Figure 3: Auto-inversion for 5 synthetic vowels. The bold grey line corresponds to the original profile. The thin black line shows the mean of the extracted profiles, which vary between the two dashed lines.

Hence, modeling a vocal tract inner contour corresponds to applying the following block-structured linear equation system of dimensions 32×7:

$$\begin{bmatrix} l_{1,1} & & & & l_{1,2} & l_{1,3} & \\ l_{2,1} & & 0 & & l_{2,2} & l_{2,3} & 0 \\ l_{3,1} & & & & l_{3,2} & l_{3,3} & \\ t_{1,1} & t_{1,2} & t_{1,3} & t_{1,4} & & & \\ t_{2,1} & t_{2,2} & t_{2,3} & t_{2,4} & & 0 & \\ \vdots & \vdots & \vdots & \vdots & & & \\ t_{25,1} & t_{25,2} & t_{25,3} & t_{25,4} & & & \\ la_{1,1} & & & & & & la_{1,2} \\ la_{2,1} & & 0 & & & & la_{2,2} \\ la_{3,1} & & & & & & la_{3,2} \\ la_{4,1} & & & & & & la_{4,2} \end{bmatrix} \begin{bmatrix} jw \\ tp \\ ts \\ tt \\ lh \\ lp \\ lx \end{bmatrix} = \begin{bmatrix} \mathrm{LIP_{pr}} \\ \mathrm{LIP_{ap}} \\ \mathrm{LIP_{wd}} \\ \mathrm{TNG}_1 \\ \mathrm{TNG}_2 \\ \vdots \\ \mathrm{TNG}_{25} \\ \mathrm{LRX}_{x_1} \\ \mathrm{LRX}_{y_1} \\ \mathrm{LRX}_{x_2} \\ \mathrm{LRX}_{y_2} \end{bmatrix} \quad (3)$$

where $l_{i,i}$, $t_{i,i}$ and $la_{i,i}$ are the factors determined for the lips, tongue and larynx respectively. The features resulting from this linear combination are projected back to the semipolar grid to obtain the vocal tract contour.

Fixed length approximation of the model - The area function, and thus the profile, must have a fixed length to permit a connection with lattice filtering. Hence, the presented system makes use of some approximations:

- the lip protrusion parameter lp and the related factors are ignored. Lips protrusion $\mathrm{LIP_{pr}}$ is not computed; it is fixed to 0.5cm.
- the lip width quantity $\mathrm{LIP_{wd}}$ is also ignored; the lip section's area is derived from the lip aperture by direct application of the $\alpha\beta$ transform.
- the larynx shape is approximated by assuming that the two larynx points LRX_{x_1,y_1} and LRX_{x_2,y_2} always move along the second grid line. The larynx height parameter lx is not used.
- a fixed area glottis section is added along the last grid line, with a fixed section of $1.5cm^2$.

The subsequent lines and columns are thus removed from (3), reducing its dimensions to 30×5. This approximation does not affect tongue shapes, but affects overall vocal tract length through blocking lip protrusion and larynx height variations. This is likely to entail some inaccuracies in the synthesis and inversion of some French vowels, such as /u/ and /y/ (Worldbet phonemic notations), for which lip protrusion plays an important role.

Inversion - The fixed length version of linear system (3) comprises 30 equations for 5 variables. Hence, inverting it, i.e. finding the values of the control parameters given a particular shape, amounts to solving an overdetermined linear problem. Solutions for such problems are available through Least-Squares solving:

$$Vw = s \quad \rightsquigarrow \quad \hat{w} = (V^T V)^{-1} V^T s \qquad (4)$$

where V is the matrix of known factors, s is the vector describing the tract shape, and w is the vector of articulatory parameters ($\hat{w}$ being its Least Squares estimate). This method finds the closest shape (in a Least Squares sense) that the linear model can produce with respect to the given shape. Hence, it performs a model-driven smoothing of the input shape.

Fortunately, in the fixed length case, the factors matrix V has full rank and is well conditioned. Numerous algorithms such as Singular Value Decomposition (SVD) and QR factorization are thus available to solve the problem [GL83]. SVD has been used in the current implementation (LAPACK routine `dgelss`), but it could be substituted with any other algorithm liable to bring more adapted or more accurate solutions.

3. ACOUSTICO-ARTICULATORY INVERSION RESULTS

Auto-inversion - As a first assessment, it is useful to verify whether information losses that occur within Least-Squares smoothing, area functions resampling and reflection coefficients estimation still allow for recovery of synthetic template tract shapes.

Hence, a set of synthetic vowels has been produced, using articulatory parameters that correspond to cardinal French vowels registered in the UPSID phonemic database. Informal listening tests ensured that the produced synthetic vowels were acceptable despite the fixed length approximation and the lossless LPC synthesizer. A more complete evaluation of the synthesis capabilities of the system should nevertheless be performed, e.g. with formant measurements and comparison with human values.

Results depicted in figure 3 show that the estimated shapes lie close to the original synthetic shapes. The observed variations result from the pulse train excitation used for synthesis.

Inversion of real speech - The system has been used to invert real speech recorded from a French male speaker in a quiet environment. Several vowel sequences and VCV sequences have been tested. Results corresponding to a "vocalic triangle" (/i e E A o u/ sequence) and to the /A b i/ sequence are given in figures 4 and 5. The system appears to locate constrictions at phonetically relevant places of articulation (e.g. front for /A/, back for /i/). Lip apertures also seem realistic. In the /A b i/ sequence, the /b/ consonantal closure appears to be detected. A spurious closure is nevertheless observed in some cases at the back of the tongue.

The obtained results are good from a qualitative point of view. Further work includes comparing them with human data to assess their accuracy.

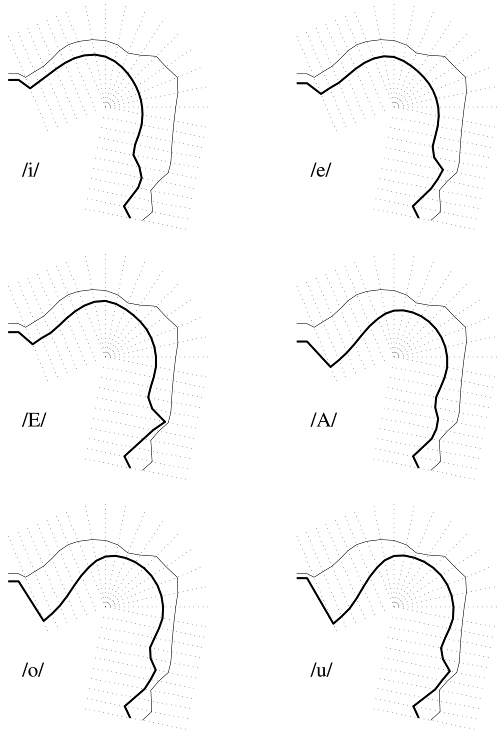

Figure 4: Inversion of human vowels.

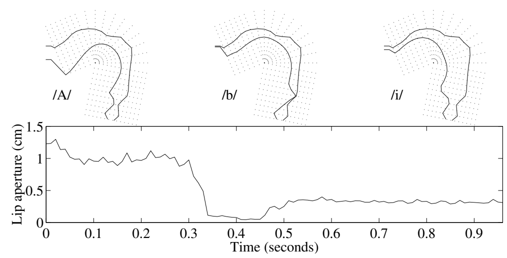

Figure 5: Inversion of /abi/.

4. ASSETS OF THE METHOD

Real-time computation - The method is analytic from end to end, thus allowing for real-time computation of articulatory features from speech.

Modularity - Any of the blocks used in the acoustico-articulatory chain (fig. 1) can be replaced with a more elaborate or more precise component, provided the replaced block admits a reciprocal. Hence, current limitations may be alleviated in future versions by using a more detailed profile shape model, better profile-to-area transformations, or acoustic estimators incorporating more elaborate relations to speech production.

Links with Digital Signal Processing (DSP) - The method creates a link between articulatory modeling and the whole gear of parametric DSP tools (all-pole spectral modeling, spectral distortion measures, parametric speech coding methods, etc.). Hence, "articulatory speech processing" might be envisioned in the long term:

- speech could be coded as low bit-rate articulatory trajectories;
- spectral estimates could be constrained by acting on articulatory trajectories (e.g., smoothing, or thresholding with reference to the human range);
- segmental speech recognition models could exploit the smoothness of the estimated articulatory features.

Further work is of course needed to determine whether such applications are viable.

ACKNOWLEDGMENTS

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

5. REFERENCES

[GL83] G.H. Golub and C.F. Van Loan. *Matrix computations.* Johns Hopkins Univ. Press, 1996 (3rd) edition, 1983.

[LS96] V. Lecuit and A. Socquet. Conséquences acoustiques du passage de la coupe sagittale a la fonction d'aire. In *Proc. XXIe JEP*, 1996.

[Mae79] S. Maeda. Un modèle articulatoire de la langue avec des composantes linéaires. In *Actes des 10èmes Journées d'Études sur la Parole*, pages 154–162, 1979.

[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.

[Sun69] J. Sundberg. On the problem of obtaining area function from lateral x-ray pictures of the vocal tract. *STL-QPSR*, (1):43–45, 1969.

[Wak79] H. Wakita. Estimation of vocal-tract shapes from acoustical analysis of the speech wave: the state of the art. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(3):281–285, 1979.