LPC MODELING WITH SPEECH PRODUCTION CONSTRAINTS

Sacha Krstulović
IDIAP C.P. 592 - CH-1920 Martigny - Switzerland
Email: sacha@idiap.ch

ABSTRACT

Despite the approximations it supposes, performing LPC-based acoustico-articulatory inversion is justified in some applicative frameworks. By illustrating this assertion through experiments aiming at incorporating speech production constraints from the DRM model and from a factor-based model into an LPC modeling scheme, we promote the use of LPC-based inversion as an interface between Production Modeling and Automatic Speech Processing methods.

1 INTRODUCTION

Speech production models usually attempt at mirroring the coarticulation processes, and rely on quantities that stay, at one level or another, proportional to measurements from human speakers. In this respect, they are ideal candidates for representing speech in Automatic Speech Processing (ASP) applications : in an Automatic Speech Recognition framework [RSS96], prior knowledge about coarticulation would help building more elaborate phoneme models and would help modeling intra-speaker variability; in a speech de-noising framework, it would help characterizing sounds not producible by humans.

But if ASP has broadly benefited from interactions with the Auditory Modeling community, e.g. through the use of the Mel scale, RASTA-PLP or more elaborate cochlear models in the feature extraction process, few concluding proposals have been made concerning the use of speech production models for speech parameterization.

Nevertheless, it is traditionally recognized that Linear Prediction Coefficients (LPC) modeling of speech is based on a production model. This model is un-specialized, in the sense that it allows the modeling of any sound, including non-speech, with an equal accuracy. As a matter of fact, few knowledge about speech production is reflected in the equations of LPC modeling : the corresponding source+filter model is completely unconstrained beyond its Auto-Regressive (AR) nature.

On the other hand, it has been demonstrated [MG76, Wak79] that the process of AR filtering was, under certain conditions, formally equivalent to acoustic filtering by lossless, rigid tubes discretized in equal-length, time-varying sections. It is interesting to note that most of today's speech production models rely upon a tube model at the articulatory/acoustic interface level (for synthesis tasks). Hence, mirroring special articulatory characteristics of these models as constraints in the LPC estimation scheme appears to be a reasonable way of using speech production modeling theories into ASP techniques.

We apply this idea to two speech production models, the Distinctive Regions and Modes (DRM) model [MCG88], and a factor-based vocal-tract sagittal cut model similar to Maeda's model [Mae79] or ICP's model [BBB+96]. After having reviewed the general framework of the AR filtering/tube filtering equivalence, we will explain how we introduced DRM-derived constraints into LPC estimation. We will expose the improvements brought by these constraints. Next, we will explain how LPC can be used as a means of inverting a factor model. Results about the extraction of sagittal cuts will then be given.

2 LPC AND ARTICULATION : GENERAL FRAMEWORK

2.1 General method

It has been shown by several authors [MG76, Wak79] that the process of Auto-Regressive digital filtering, also known as the Linear Prediction process, was analogous to acoustic filtering in discrete lossless acoustic tubes provided that :

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

If, in addition, the speech signal is pre-emphasized to compensate for spectral characteristics of the glottal excitation source and for radiation impedance at the lips, playing with this formal analogy allows to recover vocal tract area functions from the speech waveform by application of well known inverse filtering techniques [Mak77, Wak79].

Despite the mathematical elegance and computational efficiency of this method, it has found few echo in today's investigations of acoustico-articulatory inversion methods : those are mainly based on codebooks, functional approximations using neural networks, or adaptation of parameters through optimization of a synthesis model. This relative lack of success is mainly due to :

- difficulties in evaluating its accuracy : apart from comparisons with Fant's Russian vowels data and comparisons with synthetic area functions from the Ishizaka-Flanagan model, no convincing evaluation had been performed.
- the fact that the addition of losses or a nasal tract are difficult in the framework of this model.

2.2 Evaluating the method

At the time of creation of Wakita's method [Wak79], vocal tract shape measurements were sparse and computational means to exploit them were low. Twenty years after, X-ray movies or Magnetic Resonance Imagery (MRI) pictures of vocal tracts in action are more accessible (although still costly). The yet unsolved problem resides in the transformation of vocal tract sagittal cuts into area functions. Hence, a reliable method to directly measure human area functions is still unavailable.

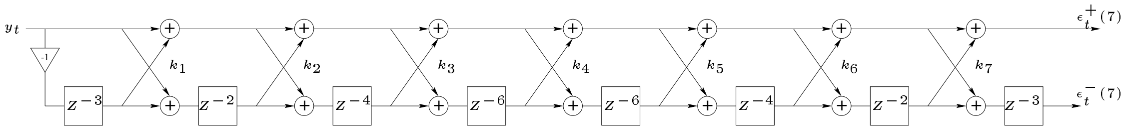

Figure 1: Lattice filter accounting for the DRM configuration.

Therefore, qualitative assessment of LPC-based acoustico-articulatory inversion is now possible (cavities at the "right" places, correct tendencies for lips or tongue movements), but providing the actual order of magnitude of the modeling error is still impossible. Hence, we can not aim at modeling perfectly the human reality (this stays an ideal ultimate goal). Rather, we attempt to match available articulatory modeling knowledge with well known signal processing methods to represent more typical human articulatory phenomena in ASP speech features. Assessment of our methods has therefore a meaning with respect to the applicative goal, more than with respect to an ideal human modeling.

2.3 Losses and nasal tract

It is often argued that the LPC/acoustic tube analogy is inherently bad as a speech production model because it does not incorporate a model of losses nor a model of the nasal tract. In principle, adding losses or a nasal tract simply amounts to changing the vocal tract's transfer function form, i.e. adding some zeros in addition to the poles (ARMA modeling instead of AR modeling). In practice, it appears that the expression of the transfer function gets far much complicated [Ole95, MG76], and the corresponding estimation process difficult to manage. Hence, the necessity for such costly refinements becomes in turn questionable :

- losses can be considered as negligible with respect to the amplitude of acoustic resonance phenomena [Wak79],
- alternately, formant frequency shifts produced by "forgetting" the losses may not necessarily be significant with respect to perception or an application in speech recognition or de-noising,
- modeling non-nasalized sounds allows to cover a part of the "speech space" which is sufficient for most ASP applications.

LPC-based acoustico-articulatory inversion does imply some approximations, but here again, approximations and performances are to be assessed with respect to the applicative goal rather than with respect to a human "reality" which we do not know how to reliably compare with.

3 CONSTRAINING ACOUSTIC FEATURES ESTIMATION WITH THE DRM MODEL

3.1 DRM inversion method

The DRM acoustic filtering process is equivalent to a lattice filtering process including articulatory constraints in the form of odd delays [Krs99] (fig. 1). Denoting by Σ_p the sum of all delays from order 1 to order (p) and applying Burg's method, a stable filter characterized by its reflection coefficients $k_{(p+1)}$ can be estimated by minimizing a mean squared prediction error $\xi(p+1)$:

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

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

Alternately, an estimator of the Itakura-Saito form can be applied :

$$k_{p+1} = \frac{-\sum_{t=\Sigma_p+1}^{N} \epsilon_t^+(p)\epsilon_{t-n_p}^-(p)}{\sqrt{\sum_{t=\Sigma_p+1}^{N}\left(\epsilon_t^+(p)\right)^2 \sum_{t=\Sigma_p+1}^{N}\left(\epsilon_{t-n_p}^-(p)\right)^2}} \quad (3)$$

This last estimator does not correspond to the minimization of an error criterion, but is based on statistical considerations [Mak77].

Hence, the DRM inversion method takes the following steps :

1. Low-pass filter speech up to 4kHz, resample to 30kHz (polyphase method) and pre-emphasize to comply with the conditions of validity of the LPC/tube equivalence.
2. Apply one of the abovementioned acoustic estimators to extract reflection coefficients k_i. (Observation window length : 25ms; window shift : 10ms.)
3. Deduce area function from reflection coefficients :

$$k_i = \frac{S_{i+1} - S_i}{S_{i+1} + S_i} \quad \Leftrightarrow \quad S_i = S_{i+1}\frac{1-k_i}{1+k_i} \quad (4)$$

   where sections are numbered from lips to glottis. A starting glottis area has to be specified : it is usually fixed to $S_P = 1.5cm^2$.

3.2 Results

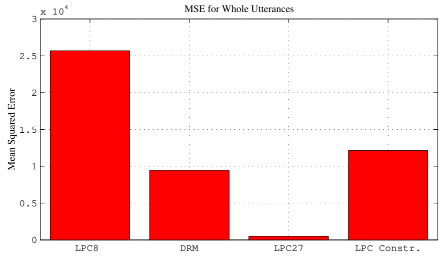

Figure 2: Comparison of Mean Squared Errors for the different tube models. Upsampling or downsampling is performed from the original 21 kHz sample frequency.

Modeling accuracy Figure 2 shows the Mean Square residual prediction Error (MSE) for an 8th order unconstrained filter, then the DRM filter, which has 8 parameters but a transfer function of order of 27, then a 27th order unconstrained filter and finally an 8 parameter, 27th order tube with different length constraints than the DRM. From this figure, it is clear that the DRM-constrained filter has better modeling performance than an unconstrained LPC model with an equal number of parameters (or equivalently a tube model with 8 equal-length sections). Of course, its performance stays lower than that of the unconstrained tube of the same LPC order. The last column indicates that the repartition of the section's length plays a role in the modeling accuracy, since a tube

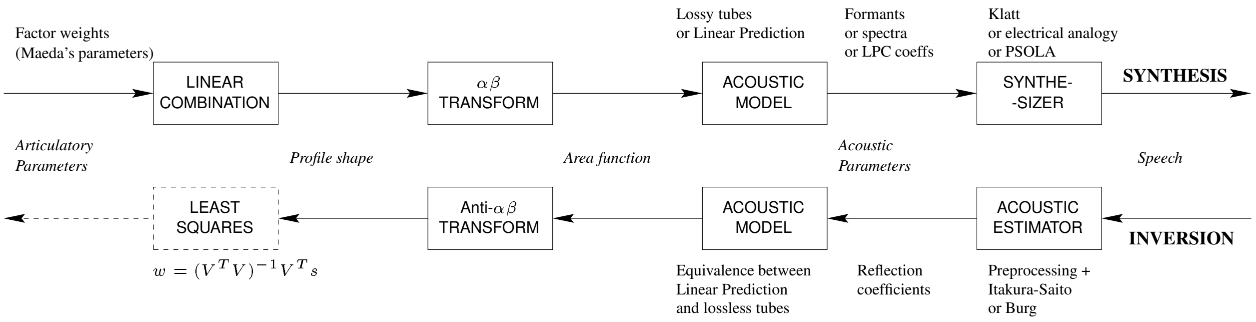

Figure 3: The acoustico-articulatory processing chain for a linear vocal tract shape model.

with the same order than the DRM but a different repartition of lengths produces a higher modeling error.

Speech recognition accuracy Table 1 shows the word error rates obtained on a medium vocabulary, speaker independent speech recognition task [Krs99]. Results indicate that Log Area Ratios (LAR) inheriting the DRM constraints perform better than LAR corresponding to an 8th order LPC model for both reflection coefficients' estimators.

Feature type (+E+Δ+$\Delta\Delta$)	WER	Relative gain
LAR	16.18%	
LAR_DRM_BURG	13.45%	16.9%
LAR_DRM_ITAKURA	12.81%	20.8%

Table 1: Recognition results obtained with the DRM-constrained log-area ratios with the two different acoustic estimators.

Comparison with X-rays Comparisons of the area functions with X-ray data has not been performed, since we don't dispose of an adequate geometric transformation between sagittal cuts and transfer functions with DRM constraints.

4 CONSTRAINING ACOUSTIC FEATURES ESTIMATION WITH A LINEAR FACTORS SHAPE MODEL

4.1 Linear model inversion method

The method, illustrated by figure 3, decomposes into the following steps :

1. same as in the case of the DRM : low-pass filter, resample and pre-emphasize.

2. Apply the (unconstrained) Itakura-Saito acoustic estimator [Mak77] to extract reflection coefficients :

$$k_i = \frac{-\sum_{t=0}^{N} \epsilon_t^+(i)\epsilon_t^-(i)}{\sqrt{\sum_{t=0}^{N}\left(\epsilon_t^+(i)\right)^2 \sum_{t=0}^{N}\left(\epsilon_t^-(i)\right)^2}} \qquad (5)$$

   with observation window length of 25ms and window shift of 10ms. Alternate estimators (e.g. Burg [Mak77] or Levinson) could also be applied.

3. Deduce area function from reflection coefficients (eq. 4).

4. Transform area function into vocal tract profiles, through application of the $\alpha\beta$-transform [HS65] or one of its more recent versions :

$$S_i = \alpha(i) d_i^{\beta(i)} \quad \Leftrightarrow \quad d_i = \left(\frac{S_i}{\alpha(i)}\right)^{\frac{1}{\beta(i)}} \qquad (6)$$

A yet unimplemented further step would be :

5. Decompose the obtained shapes into a linear components basis similar to Maeda's shape basis. The main difference with Maeda's model comes from the fact that the LPC analogy does not allow a time varying vocal tract length.

   Since the corresponding linear system comprises 30 equations for 7 unknowns (provided that 7 factors are used), the solution can be obtained through Least-Squares solving :

$$s = wV \quad \leadsto \quad \hat{w} = (V^T V)^{-1} V^T s \qquad (7)$$

   where V is the matrix of known factors, s is the vector describing the tract shape, and w is the vector of factors' weights ($\hat{w}$ being its least squares estimate).

4.2 Results

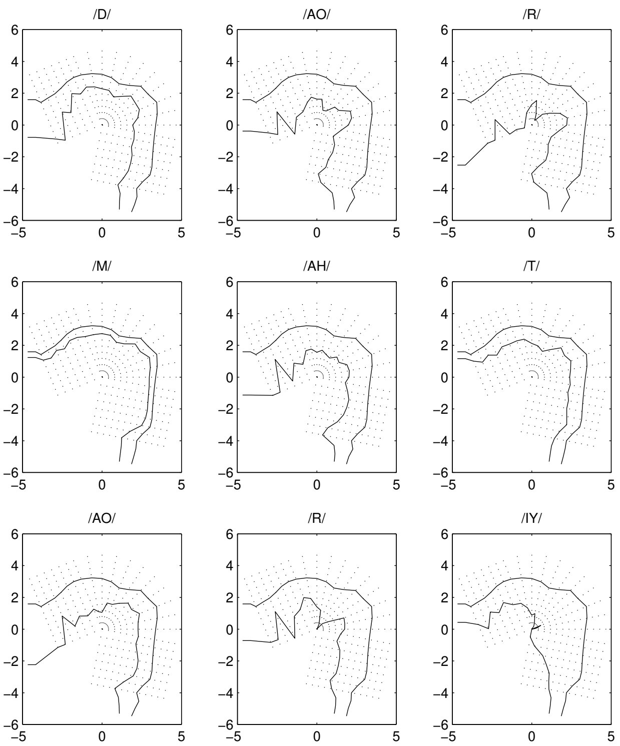

Figure 4: Pronunciation of the English word "dormitory". Example of shapes automatically extracted from a sound file of the XRDB database (Itakura-Saito acoustic estimator).

Qualitative assessment Figure 4 shows a sample of the sequence of shapes obtained on a sound file of the University of Wisconsin's XRDB database after step 4 of the method. From this example, qualitative considerations can be emitted : cavities for /AO/, /AH/ and /IY/ are globally at the "right place" (back for /IY/, front for /AO/); incisors seem represented on the 3rd grid line for all the phonemes; the /T/ and /D/ plosives actually imply that the tongue comes close to the palate (this is followed by a release not represented on the figure); there is a lip closure for the /M/; the particular shape of the /R/ can be interpreted as a retroflexion of the tongue; the premises of the /R/'s lingual configuration are visible in the preceding /AO/, which means that *coarticulation phenomena become observable*. These observations can be generalized to the totality of the spoken sequence, which is in fact 7 words long, as well as to other sequences of the database. Moreover, we have observed that the obtained trajectories are smooth, i.e. there are no discontinuous jumps from one shape to the following (see for instance lower lip trajectory on fig. 5).

These observations have of course to be relativized. For the /D/ and the /T/, actual contact is not observed because the analysis scale is too large (25ms windows shifted by 10ms are unable to capture it) : adaptation of these scales (with a long mode and a short mode, or a pitch synchronous analysis) could alleviate this problem. Next, the shape of the /R/ is made shocking by its lack of smoothness, due to the effects of the application of the anti-$\alpha\beta$ transform on the grid lines only : this can be corrected by changing the mapping function into a more "smoothing" one. Finally, for the /M/, only the lips should close, not the whole tract : we do not have yet a solution to this problem.

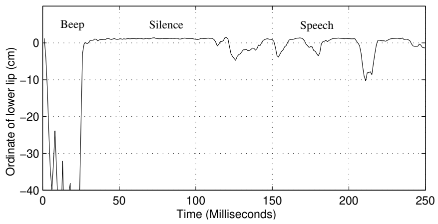

Figure 5: Ordinate of the trajectory of the "lower lip". The "beep" at the beginning of the recording gives rise to values clearly out of the range of those of a human speaker.

Further observations and forecasted applicative goals When superimposing phonetic alignments, that were automatically generated from forced alignment of acoustic Hidden Markov Models (HMMs), with the sequences of extracted shapes, we have observed that the phoneme boundaries in the acoustic domain were becoming questionable in the sagittal shapes domain. Hence, it is reasonable to forecast that embedded training of HMMs, together with their alignment in the decoding pass, will converge towards a different solution than the one based on acoustic features. Whether this solution gives better or worse recognition scores is currently investigated.

Figure 5 shows that the "beep" present at the beginning of the recordings was giving rise to values clearly out of the human range. Hence, the "beep" zone could be detected by simple thresholding of the obtained values, which would represent a simple denoising method. A deeper knowledge of the system's behavior could allow to extend this type of application.

5 CONCLUSION

We have shown that the incorporation of DRM-derived constraints into the process of LPC modeling was bringing improvements to the modeling accuracy and to a speech recognition task. Alternately, obtaining vocal tract profile shapes through LPC, with the ultimate goal of inverting factor-based vocal tract models, shows an opening towards the use of speech production paradigms into speech recognition and speech denoising applications. Hence, despite the approximations it supposes, LPC-based acoustico-articulatory inversion is a good candidate as an interface between speech production knowledge and Automatic Speech Processing applications.

ACKNOWLEDGMENTS

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

REFERENCES

[BBB+96] D. Beautemps, P. Badin, G. Bailly, A. Galván, and R. Laboissière. Evaluation of an articulatory acoustic model based on a reference subject. In *1st ESCA Tutorial and Research Workshop on Speech Production Modeling*, May 20-24 1996.

[HS65] J.M. Heinz and K.N. Stevens. On the relations between lateral cineradiographs, area functions, and acoustic spectra of speech. In *Proc. 5th Int. Congress of Acoustics*, volume A44, 1965.

[Krs99] S. Krstulović. LPC-based inversion of the DRM articulatory model. In *Proc. Eurospeech'99*, 1999.

[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.

[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.

[Ole95] M. Olesen. *A speech production model including the nasal cavity*. PhD thesis, Dept. of Communication Technology, Inst. of Electronic Systems, Aalborg Univ., Denmark, October 1995.

[RSS96] R.C. Rose, J. Schroeter, and M.M. Sondhi. The potential role of speech production models in automatic speech recognition. *J. Acoust. Soc. Am.*, 99(3), March 1996.

[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.