NEW PERSPECTIVES ON LINEAR-PREDICTION MODELLING OF THE VOCAL-TRACT: UNIQUENESS, FORMANT-DEPENDENCE AND SHAPE PARAMETERISATION

Parham Mokhtari[†] and Frantz Clermont[‡]
[†]Electrotechnical Laboratory (ETL), Japan.
[‡]University College, University of New South Wales (UC-UNSW), Australia.

It is well known that the linear-prediction (LP) model yields an inherently unique estimate of the vocal-tract (VT) shape from information contained only in the acoustic speech signal. However, the uniqueness property of the LP-VT model is understood at best incompletely, resulting in a perceived lack of confidence and a popular trend towards more sophisticated VT models and methods of acoustic-to-articulatory mapping. This paper contributes a better understanding of the LP-VT model's uniqueness property, by revealing for the first time, the formant frequency- and bandwidth-dependence of LP-derived VT-shapes. Our results thereby (i) provide a shape-related explanation of the uniqueness property of the LP-VT model, and (ii) suggest a new, acoustically-relevant parameterisation of VT-shapes.

INTRODUCTION

Since the pioneering works of Atal (1970) and Wakita (1972), it is well known that the shape of the vocal-tract (VT) during the production of vocalic speech sounds can be estimated by linear-prediction (LP) analysis of the speech waveform. Most importantly, the LP method of mapping from the acoustic to the articulatory domain overcomes the difficult problem of nonuniqueness (e.g., Atal et al., 1978; Schroeter and Sondhi, 1994), and thereby yields an inherently unique estimate of the VT-shape from information contained only in the acoustic speech signal. Indeed, that uniqueness property is specific to the LP-VT model, which is physically a concatenation of lossless, cylindrical acoustic tubes, with a single, resistive (or frequency-independent) source of acoustic energy loss at the glottal end (Wakita, 1972, 1973). However, the acoustic-articulatory relations on which the LP-VT model's uniqueness property is founded, appear to be incompletely understood; and partly as a result, the LP method of inversion has been largely neglected during the past two decades, in favour of more sophisticated and presumably more realistic modelling approaches.

By contrast with the incomplete understanding of the acoustic-articulatory properties of the LP-VT model, the relations between the formant (or acoustic resonance) frequencies and the shape of a *completely lossless* VT have long been known. Indeed, the seminal work of Mermelstein and Schroeder (1965), later expanded by each author separately (Schroeder, 1967; Mermelstein, 1967), showed that each odd-indexed coefficient a_{2n-1} of the Fourier cosine-series of the logarithmic area-function (the cross-sectional area of the VT airway as a function of the distance from the glottis to the lips) is related *uniquely* and *quasi-linearly* with the corresponding formant frequency F_n, according to the following, approximate relation:

$$a_{2n-1} \approx -2\frac{(F_n - F_{n0})}{F_{n0}}, \qquad (1)$$

where $F_{n0} = (2n-1)c/4L$ is the n^{th} formant frequency of a uniform area-function of the same length L, and $c = 35300$ cm/sec is the speed of sound in the VT airway. That remarkable result yielded the following, acoustically-meaningful parameterisation of the VT area-function $A(x)$:

$$\ln \hat{A}(x) = \ln A_0 + \sum_{n=1}^{N} a_{2n-1} \cos\big((2n-1)\pi x / L\big), \qquad (2)$$

where x is the distance along the length of the VT airway from the glottis to the lips, A_0 is an (acoustically inconsequential) area scaling factor, and N specifies the number of terms retained in the cosine series and hence the smoothness of the bandlimited area-function $\hat{A}(x)$. Apart from the convenience of a direct relation between VT-shape and acoustic parameters, the Schroeder-Mermelstein (SM) model embodied in Equations (1) and (2) provides an unprecedentedly insightful, shape-related explanation of the inherent nonuniqueness of the completely lossless VT model: for a given VT-length L, each formant frequency can be controlled uniquely by a corresponding, *antisymmetric* component of the VT-shape, while the *symmetric* shape components (or the even-indexed coefficients a_{2n} of the cosine series) have little or no acoustic consequence and therefore remain undetermined given only the information contained in the acoustic speech signal. In terms of the poles and zeros of the lip-impedance function (the driving point impedance of the VT as seen from the lips), that nonuniqueness property can be restated in the following way: while the zeros (the formant frequencies) can be measured from the acoustic speech waveform, measurement of the poles requires a special apparatus known as the lip-impedance tube (Schroeder, 1967; Gopinath and Sondhi, 1970), without which there is absent one half of the information necessary to determine a unique VT-shape.

Indeed, upon mathematically reducing the LP-VT to a completely lossless model (by effectively closing the glottis), Wakita and Gray (1975) showed that the poles and zeros of the lip-impedance function are real-valued and independent, such that both sets are needed to determine the VT-shape uniquely; by contrast, they showed that reinstatement of the finite glottal resistance of the LP-VT model renders those zeros and poles both complex-valued (i.e., having both a frequency and a bandwidth) and inter-dependent, such that only one set (preferably the acoustically-measurable zeros) suffices to uniquely determine the VT-shape. Thus, as Atal (1970) had previously foreshadowed, the LP-VT model's uniqueness property stems from its use of both the formant *frequencies* and *bandwidths*. In view of the insightful properties of the SM model reviewed earlier, two intriguing questions then arise: (i) whether, as in the completely lossless VT model, the formant frequencies are related uniquely and quasi-linearly with the antisymmetric shape components of LP-derived area-functions; and if so, (ii) whether the formant bandwidths are at all related to the symmetric shape-components which are somehow determined uniquely by the LP method of inversion and which, by contrast, remain undetermined in the completely lossless VT model.

In the remainder of this paper we seek answers to those fundamental questions, by empirical exploration of the formant-dependence of LP-derived VT-shapes. In the next section we explore the formant *frequency*-dependence of those shapes, and thereby make a direct comparison with the SM model. We then explore the formant *bandwidth*-dependence of LP-derived area-functions, with the aim of identifying those shape-components which the LP model is apparently able to resolve in its unique determination of the VT geometry. Our results lead to a new, acoustically-relevant parameterisation of VT-shapes, some implications of which are discussed in the concluding section.

FORMANT FREQUENCY-DEPENDENCE OF LP-DERIVED VT-SHAPES

In order to address the first of the two questions raised in the Introduction, we adopt the well-known method of articulatory-acoustic nomograms. In particular, starting with a uniform VT-shape ($a_{2n-1}=0$, $A_0=1$) of length L=17.65 cm, each of the SM model's parameters a_{2n-1}, $n=\{1,2,3\}$ is in turn perturbed in steps of 0.05 across the range $[-0.8,+0.8]$. At each step, an 8-section area-function is generated using Equation (2), and the LP model is used (with a nominal value of 0.7 for the glottal reflection coefficient) to synthesise the first three formants.

The formant frequency-nomograms thus obtained are shown in Figure 1, which at once reveals a negatively-sloped, quasi-linear and unique relation between each parameter a_{2n-1} and the corresponding formant frequency F_n. Indeed, these results confirm that the partially-lossy, discrete-sectioned LP-VT model shares with the completely lossless, smooth VT model, the distinctive relation between each formant frequency and the corresponding, *antisymmetric* component of the VT-shape. The mutual congruence in the basic acoustic properties of the LP and the SM models thereby confirmed, we now turn to the question of whether the formant bandwidths play a role in the LP model's unique determination of the *symmetric* components of VT-shapes.

FORMANT BANDWIDTH-DEPENDENCE OF LP-DERIVED VT-SHAPES

As clearly foreshadowed in the literature reviewed earlier, the formant bandwidths do play an important role in the LP inversion method's uniqueness property. However, apart from theoretical or mathematical treatments of the LP-VT model (e.g., Wakita and Gray, 1975), it has never been elucidated just how the bandwidths contribute to determining a unique VT-shape. In the next section we therefore explore the bandwidth-dependence of LP-derived VT-shapes.

Functional form of bandwidth-dependence

Without first venturing a hypothesis on the likely form of the VT-shape correlates of distinctive changes in the formant bandwidths, we adopt an exploratory approach whereby the LP method of inversion (i.e.,

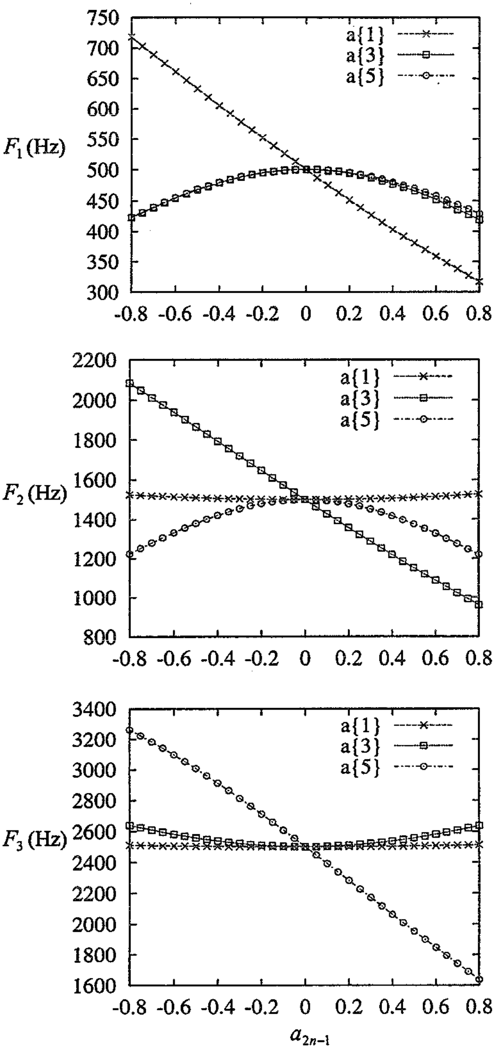

Figure 1. Formant frequency nomograms, generated by perturbing a uniform, 8-section LP area-function according to each of the first three parameters of the SM model. (Refer to text for details.)

the well-known algorithms given in Markel and Gray, 1976, for transforming a set of formants to an LP area-function) is used to obtain VT-shapes corresponding to perturbations in a single bandwidth at a time. Starting with a neutral formant-pattern given by $F_n=(2n-1)500$ Hz and $B_n=100$ Hz for $n=\{1,...,7\}$, each of the first three bandwidths in turn is perturbed in two sets of inversely-proportional steps above and below its neutral value (i.e., $B_n=\{50,80,100,125,200\}$ Hz), and at each step the LP method of inversion is used to convert the formant parameters to an area-function (of length $L=17.65$ cm). As our aim is to discern an underlying pattern in the VT-shapes thus obtained, the first 7 formants are specified in order to secure a visually satisfactory spatial resolution of 14 discrete sections per area-function.

The resulting, LP-derived (hence discrete-sectioned) area-functions are shown in Figure 2. It is remarkably apparent from the superimposed area-functions shown in each of the three panels, that perturbation of the n^{th} bandwidth induces a quasi-*sinusoidal* perturbation of the LP-derived VT-shape, with $2n-1$ half-cycles within the length of the VT from the glottis to the lips. Those implied sinusoids are shown by the dashed curves in Figure 2, superimposed with the groups of LP-derived area-functions in order to emphasise the apparent functional form of the bandwidth-induced, VT-shape perturbations.

Our results in Figure 2 therefore suggest that when a completely lossless VT model is augmented with a resistive termination at the glottal end (as in Wakita's formulation of the LP model), *symmetric* (sinusoidal) perturbations of the uniform area-function are associated with distinctive variations in the formant *bandwidths*. In elucidating the bandwidth-dependence of LP-derived VT-shapes, we have thus revealed the functional form of the acoustic-articulatory relations which the LP-VT model uses in determining a unique VT-shape from a set of formant frequencies and bandwidths. In the next section we take these results one step further and propose a new, acoustically-meaningful parameterisation of VT-shapes.

Acoustically-meaningful parameterisation of VT-shapes

Indeed, our empirical exploration of the bandwidth-dependence of LP-derived VT-shapes suggests a functional form of the symmetric shape perturbations, specifically in terms of the odd-indexed coefficients of the Fourier *sine* (rather than the SM model's original implication of the Fourier *cosine*)

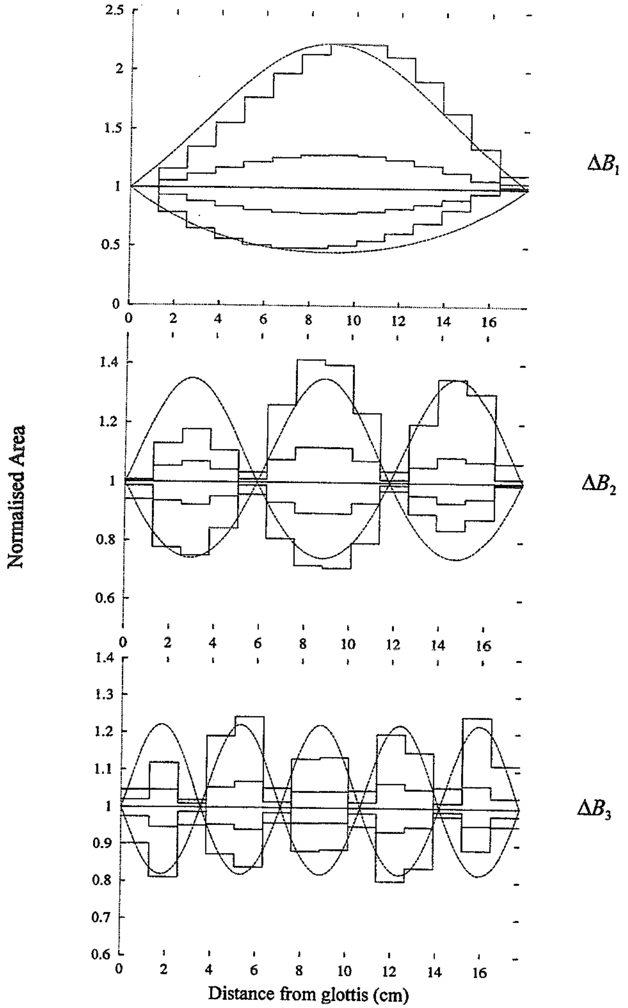

Figure 2. Bandwidth-dependence of LP-derived VT-shapes. (Refer to text for details.)

series of the logarithmic area-function. We are therefore compelled to extend the SM model's parameterisation in Equation (2), in order to account for the unique, acoustic-articulatory properties of the LP-VT model, as follows:

$$\ln\hat{A}(x)=\ln A_0+\sum_{n=1}^{N}a_{2n-1}\cos((2n-1)\pi x/L)+\sum_{n=1}^{N}b_{2n-1}\sin((2n-1)\pi x/L). \qquad (3)$$

As the new parameters b_{2n-1} form an orthogonal set with the original parameters a_{2n-1}, they together provide a mathematically complete and minimal description of the VT-shape. More importantly in the framework of the LP-VT model, those sets of parameters are distinctively related with the formant bandwidths and frequencies, and thus provide an acoustically-relevant parameterisation of VT-shapes. Equation (3) next allows us to provide one further test of our LP-related extension of the SM model.

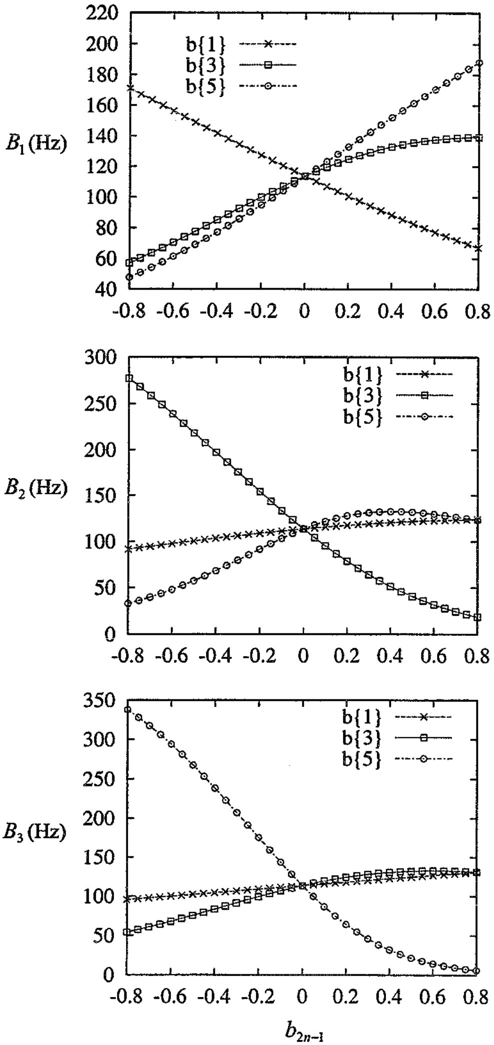

Figure 3. Formant bandwidth nomograms, generated by perturbing a uniform, 8-section LP area-function according to each of the first three, new parameters of the extended SM model. (Refer to text for details.)

Bandwidth-dependence by nomograms

Similarly to our earlier illustration of the formant frequency-dependence of LP-derived VT-shapes, we may now test the bandwidth-dependence of those shapes by way of nomograms. Starting with a uniform VT-shape ($a_{2n-1}=0$, $b_{2n-1}=0$, $A_0=1$) of length $L=17.65$ cm, each of the new parameters b_{2n-1}, $n=\{1,2,3\}$ is in turn perturbed in steps of 0.05 across the range $[-0.8,+0.8]$. At each step, an 8-section area-function is generated using Equation (3), and the LP model is used (with the same, nominal value of 0.7 for the glottal reflection coefficient) to synthesise the first three formants.

The formant bandwidth-nomograms thus obtained are shown in Figure 3, which indeed reveals a quasi-linear and unique relation between each parameter b_{2n-1} and the corresponding formant bandwidth B_n. While certain other changes in bandwidth-value are comparable in magnitude (particularly those for B_1), a unique or distinctive relation between b_{2n-1} and B_n is implied by the single, negatively-sloped nomogram shown in each of the three panels in Figure 3. Indeed, those results yield the following, bandwidth-related homologue to Equation (1):

$$b_{2n-1} \approx -2\frac{(B_n-\overline{B})}{(2n-1)\overline{B}}, \qquad (4)$$

where $\overline{B}$ is the mean of the N bandwidths of a $2N$-sectioned LP area-function, and is determined completely by the glottal reflection coefficient (Kasuya & Wakita, 1979; Mokhtari, 1998). The appropriateness of our new parameterisation in Equation (3) is thereby confirmed in the context of the LP-VT model, which clearly maps the formant bandwidths to the symmetric, sinusoidal components of VT-shapes.

SUMMARY & CONCLUDING DISCUSSION

In this paper we have adopted an empirical approach to explore the formant-dependence of LP-derived VT-shapes. We have first confirmed that the LP-VT model shares with the completely lossless VT model, the distinctive relation between each formant frequency and the corresponding, antisymmetric (cosine) component of the VT-shape. We then used the LP method of inversion to reveal that perturbations in each formant bandwidth induce symmetric (sinusoidal) perturbations in the shape of a uniform LP area-function, and bandwidth nomograms subsequently confirmed that in the LP-VT model there is indeed a distinctive relation between each formant bandwidth and the corresponding, symmetric (sine) component of the VT-shape.

Our empirical results have led to a new, acoustically-relevant parameterisation of VT-shapes which augments the odd-indexed coefficients of the Fourier cosine-series originally proposed by Mermelstein and Schroeder (1965), with the odd-indexed coefficients of the Fourier sine-series. Those two sets of parameters provide a mathematically complete and compact representation of the smoothed (or bandlimited) VT-shape, thus overcoming the limitations imposed by the LP-VT model's coarse, step-wise representation of the area-function (Mokhtari, 1998). While this new parameterisation is clearly significant in the context of the LP-VT model, it also raises fundamental questions regarding the heretofore disregarded role of symmetric VT-shape components.

That the LP-VT model effectively uses the formant bandwidths to uniquely determine the symmetric components of estimated VT-shapes, is indeed of significance in a broader context. On the one hand, the LP-VT is a simplified model of the human vocal-tract in which the only source of acoustic energy loss retained is at the glottis; the distinctive relations with the symmetric components of the VT-shape therefore arise from only the glottal component of the true (or acoustically measured) bandwidths. On the other hand, from an auditory-perceptual point of view the bandwidths have long been known to play a much less important role than the formant frequencies in the acoustic-phonetic characterisation of speech sounds (e.g., Klatt, 1982). However, it has also long been known that speech perception phenomena do not necessarily justify or adequately explain speech production phenomena. Indeed, our new results challenge the long-held inferior role of the formant bandwidths, and suggest that at least from an articulatory point of view, the glottal component of bandwidths may play an important role in the characterisation of vocalic speech sounds. Further investigations by VT-modelling and by simultaneous and direct measurements of the acoustics and articulation of speech are called for, in order to shed light on the implications of these findings.

ACKNOWLEDGEMENTS

We gratefully thank Drs David J Broad and Michael Barlow for their invariable support and encouragement.

REFERENCES

Atal, B. S. (1970). "Determination of the Vocal-Tract Shape Directly from the Speech Wave", *J. Acoust. Soc. Am.*, Vol. 47, S65.

Atal, B. S., Chang, J. J., Mathews, M. V. and Tukey, J. W. (1978). "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", *J. Acoust. Soc. Am.*, Vol. 63, 1535-1555.

Gopinath, B. and Sondhi, M. M. (1970). "Determination of the Shape of the Human Vocal Tract from Acoustical Measurements", *The Bell System Technical Journal*, Vol. 49, 1195-1214.

Kasuya, H. and Wakita, H. (1979). "An Approach to Segmenting Speech into Vowel- and Nonvowel-Like Intervals", *IEEE Trans. Acoust., Speech, and Sig. Process.*, Vol. 27, 319-327.

Klatt, D. H. (1982). "Prediction of perceived phonetic distance from critical-band spectra: A first step", *Proc. Int. Conf. on Acoust., Speech, and Sig. Process.*, 1278-1281.

Mermelstein, P. (1967). "Determination of the vocal-tract shape from measured formant frequencies", *J. Acoust. Soc. Am.*, Vol. 41, 1283-1294.

Mermelstein, P., and Schroeder, M. R. (1965). "Determination of smoothed cross-sectional area functions of the vocal tract from formant frequencies", *Proc. 5th Int. Congr. Acoust.*, Liège, Paper A24.

Mokhtari, P. (1998). "An acoustic-phonetic and articulatory study of speech-speaker dichotomy", Unpublished PhD Thesis, The University of New South Wales, Australia.

Schroeder, M. R. (1967). "Determination of the geometry of the human vocal tract by acoustic measurements", *J. Acoust. Soc. Am.*, Vol. 41, 1002-1010.

Schroeter, J. and Sondhi, M. M. (1994). "Techniques for Estimating Vocal-Tract Shapes from the Speech Signal", *IEEE Trans. Speech and Audio Process.*, Vol. 2, 133-150.

Wakita, H. (1972). "Estimation of the vocal tract shape by optimal inverse filtering and acoustic/articulatory conversion methods", Speech Communications Research Laboratory, Inc., Santa Barbara, California, USA, Monograph No. 9.

Wakita, H. (1973). "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms", *IEEE Trans. Audio and Electroacoustics*, AU-21, 417-427.

Wakita, H. and Gray, A. H. (1975). "Numerical Determination of the Lip Impedance and Vocal Tract Area Functions", *IEEE Trans. Acoust., Speech, and Sig. Process.*, ASSP-23, 574-580.