Chapter 5
Vocal-Tract Shape Parameterisation and Estimation

5.1 Introduction

The spectral manifestations of the interactions between phonetic and speaker-specific attributes of steady-state vowels of spoken Australian and American English, were unveiled in the preceding chapter, by way of classification experiments based upon the linear-prediction (LP) cepstrum. The flexibility afforded by our parametric cepstral distance measure (PCD, derived in Section 4.2.3.1) indeed complemented the whole-spectrum representation of the LP cepstrum, and allowed direct examination of the spectral ranges which predominantly contain either phonetic or speaker-specific influences. A more detailed examination of the vowel-speaker dichotomy was then performed by having recourse to the spectral resonance (or formant) frequencies. Although the task of formant-estimation is considerably more problematic than that of obtaining the LP cepstrum, the interpretive superiority of the formants led to a more revealing, acoustic-phonetic explanation of the dichotomy. As set out in our Introduction (Chapter 1), we now seek an even more fundamental interpretation of the vowel-speaker interactions which have thus far been elucidated only in the acoustic domain, by extending our investigations into the domain of speech production. A first step towards this end is taken here by considering the intertwined problems of vocal-tract shape estimation and parameterisation.

As suggested in our literature review (Chapter 2), difficulties in acquiring directly-measured physiological data have to a large extent prohibited a detailed, or a large-scale investigation of the articulatory sources of speaker variability. Acoustic-to-articulatory mapping, on the other hand, is potentially a more efficient and practical method, which may therefore facilitate investigations of the articulatory correlates of the many types of
variability in the acoustic speech signal. Sadly, the prospect of utilizing estimated (rather than directly measured) articulatory data seems to have suffered from want of a “perfect” solution to the so-called inverse problem, which should be capable of recovering, with a large degree of confidence, the actual vocal tract configurations of any given speaker. Our review of the relevant literature (in Section 2.4.3.2) has already underscored the crucial role of both the articulatory parameterisation and the vocal-tract acoustic model in determining the intrinsic nonuniqueness of the inverse mapping. In our pursuit of an articulatory explanation of the phenomenon of dichotomy, we therefore seek a method of acoustic-to-articulatory mapping which employs a vocal-tract acoustic model with the least degree of intrinsic nonuniqueness.

In addition, we seek a vocal-tract shape parameterisation which not only inherits the uniqueness properties of the chosen vocal-tract acoustic model, but also permits a direct relation with the acoustic parameters used in the previous section to provide an acoustic-phonetic explanation of the dichotomy. In this vein, it is hardly surprising that the formants, which were ideally suited to providing an acoustic-phonetic explanation, are also the acoustic parameters which have been traditionally, and remain to this day, the most directly associated with the physical properties of the vocal-tract. It follows that, in order to obtain an articulatory interpretation of the dichotomy analogous to the acoustic-phonetic explanation offered in Chapter 4, it would be of considerable benefit to adopt a vocal-tract shape parameterisation which is related as directly as possible to the formants.

Our aim in this chapter is therefore to develop an approach to the estimation of vocal-tract area-functions from the acoustic speech waveform of non-nasalised vowels, which satisfies two main criteria: (1) by necessity, the model adopted for the inverse mapping must be capable of yielding inherently unique vocal-tract shapes; and (2) the area-function parameterisation should be expressible directly as a function of the acoustic resonances of the vocal-tract.

In Section 5.2 we address these two issues of resonance-based parameterisation and uniqueness, and thereby provide a bipartite rationale for our proposed approach. Theoretical and empirical results are brought to bear in Section 5.3, on the problem of determining the resonance-based parameters of unique, LP-derived area-functions. In
Section 5.4 we then describe our hybrid method of area-function estimation, and use it to evaluate our proposed method of vocal-tract shape parameterisation. In Section 5.5 we first introduce a new method of quantifying inter-repetition variability amongst vocal-tract shapes, which then allows an evaluation of our hybrid method of inversion. We conclude in Section 5.6 with a summary of the contributions arising from our work in vocal-tract shape estimation and parameterisation, and with a discussion of its significance in the context of our articulatory explanation of the dichotomy which is to follow in Chapter 6.

5.2 Rationale

Our criteria for an appropriate method of acoustic-to-articulatory mapping, as stated above, incite a twofold rationale concerning parameterisation and uniqueness of estimated vocal-tract shapes, which we therefore consider, respectively, in the next two sections.

5.2.1 Resonance-Based Parameterisation of Area-Functions

The seminal work of Schroeder and Mermelstein (1965), subsequently expanded by each author separately (Schroeder, 1967; Mermelstein, 1967), is an acoustically-motivated and theoretically-derived parameterisation of the vocal-tract area-function. In contrast with the many physiologically-based models of the vocal-tract which have appeared in the literature, the parameterisation derived by these authors is rooted in the acoustic, and more specifically in the resonance, properties of the vocal-tract. It is therefore re-examined here in some detail, with a view to highlighting its strengths and limitations, and placing it in the context of our forthcoming, articulatory investigation of the vowel-speaker dichotomy.

In order to better appreciate the implications of the modelling paradigm pioneered by those two authors, we have included in Appendix B, a mathematical re-derivation of what we hereafter will refer to as the Schroeder-Mermelstein (SM) model. Indeed, as re-derived therein, the complete model is summarised in the following two equations:

\[ \ln A(x) = \ln A_0 + \sum_{m=1}^{M} a_m \cos \left( \frac{m \pi x}{L} \right), \]
Figure 5.1: Nomograms depicting changes in the first four formant frequencies, as a function of perturbations in each of the first six, vocal-tract shape-parameters of the SM model. Each area-function is of length $L=17.65$ cm, and the area scaling factor is $A_c=1.0$, consistently with the assumptions of the SM model, the formant frequencies are synthesised using a completely lossless vocal-tract acoustic model (see Appendix C).
which defines the parameterisation of the vocal-tract (logarithmic) area-function $A(x)$ of length $L$, in terms of the coefficients $a_m$ of its Fourier cosine series; and

$$\frac{\delta F_n}{F_n} = -\frac{1}{2}a_{2n-1}, \quad (5.2)$$

which defines the theoretical relation between the acoustic- and the articulatory-domain parameters. In particular, as illustrated in the nomograms on the left in Figure 5.1, antisymmetric perturbation of the area-function of a uniform, completely lossless acoustic tube according to each of the odd-indexed shape parameters, does induce a quasi-linear perturbation of a unique formant frequency about its neutral value. By contrast, as shown in the nomograms on the right in Figure 5.1, perturbation of an even-indexed parameter $a_{2n}$ has no first-order\(^1\) effect on the formant frequencies. The SM model thereby establishes an acoustically-important distinction between symmetric and antisymmetric perturbations of a uniform area-function (where the symmetry is defined with respect to the mid-point along the length of the area-function).

Not long after its inception, the SM model’s potential in yielding articulatory insights into acoustic speech phenomena was demonstrated by Broad (1972, p.418), who combined Equations 5.1 and 5.2, then differentiated the result with respect to time to obtain a “measure of the average acoustically detectable movement throughout the vocal tract.” The coarse, phonetic segmentation (of an all-sonorant utterance) thus afforded, was then refined by integrating the expression only across a selected length of the vocal-tract, thereby obtaining a segmentation based on the acoustically-inferred velocity profile of, for example, the tongue dorsum. The SM model was also advocated by Broad and Wakita (1977), who showed that it aptly predicts the $F_2$-dependence of vocal-tract shapes, in terms of the distinction between front and back places of constriction; similarly, the top-left panel in Figure 5.1 illustrates the $F_1$-dependence, which is aptly predicted in terms of the open versus closed distinction.

However, despite the interpretive potential of the SM model, it is known to suffer the limitation of nonuniqueness. In particular, while the antisymmetric-shape parameters $a_{2n-1}$ are each associated with a unique formant frequency $F_n$, the

\(^1\) By “first-order”, we refer not only to the magnitude of the induced perturbation, but also to its monotonicity (or quasi-linearity).
symmetric-shape parameters $a_{2n}$ remain ambiguous in the inverse mapping. In this vein, Borg (1946) is credited (by Schroeder, 1967; Mermelstein, 1967; Schroeter and Sondhi, 1994) to have proven that not a single, but a doubly-infinite set of eigenvalues, which correspond to two different sets of boundary conditions, is both necessary and sufficient to uniquely determine the state of a completely lossless resonant system. The lip impedance function, which is defined as the driving-point impedance of the vocal-tract as seen from the lips, is particularly useful in this context, because it embodies knowledge of both sets of eigenvalues; in terms of the complex frequency variable $s = j\omega$, it is expressed as follows:

$$Z_\text{in}(s) = \frac{\prod_{n=1}^{\infty} (s - \omega_n)}{\prod_{n=1}^{\infty} (s - \omega_n^{(c-c)})},$$

(5.3)

where the zeros correspond to the well-known formants, and the poles of the lip impedance function correspond to the resonances under “closed-closed” (c-c) boundary conditions (implying an infinite acoustic impedance at both ends of the acoustic tube). Analogously to Equation B.13 (in Appendix B), that second set of eigenvalues is given by the following expression:

$$\lambda_n^{(c-c)} = \frac{\omega_n^{(c-c)}^2}{c^2} = \left(\frac{n\pi}{L}\right)^2, \quad n = 1, 2, 3, \ldots,$$

(5.4)

and the “c-c” resonance frequencies themselves interleave with the formant frequencies, as follows:

$$\omega_1 < \omega_1^{(c-c)} < \omega_2 < \omega_2^{(c-c)} < \omega_3 < \ldots.$$

(5.5)

Both Schroeder (1967) and Mermelstein (1967) have pointed out that the poles of the lip impedance function are related to the even-indexed parameters $a_{2n}$ of the SM model in the same way that the formants are related to the odd-indexed parameters (in Equation 5.2). However, those poles can only be measured using a special apparatus known as a “lip impedance tube” (Schroeder, 1967; Gopinath and Sondhi, 1970) which has a mouthpiece sealed to the speaker’s lips, and which requires articulation of sustained vowels without phonation. The nonuniqueness problem then stems from the fact that the acoustic speech signal contains resonance information pertaining to only one of the required, two sets of boundary conditions — namely, the formants (or the
zeros of the lip impedance function, but not its poles); hence, the symmetric components of vocal-tract shapes remain undetermined in the inverse mapping, unless additional constraints are imposed.

Indeed, Mermelstein (1967) imposed the constraint of antisymmetry (which is, after all, a corollary of the SM model itself), simply by setting the undetermined, even-indexed parameters to zero. While this turned out to be a reasonable constraint for a number of vocalic configurations, it led to a poor re-estimate of the X-ray measured area-function (Fant, 1960) for the Russian vowel /u/ in particular, which has a strong symmetric component. More recently, Yehia and Itakura (1994, 1996) proposed to estimate the values of the first eight parameters of the SM model, using only the first three formant frequencies; in particular, the ambiguity associated with the values of the five shape-parameters $a_2$, $a_4$, $a_6$, $a_7$, and $a_8$, was resolved by a combination of static and dynamic constraints on the estimated shapes themselves. Although, as reviewed much earlier (in Chapter 2), the nonuniqueness problem is exacerbated when the number of vocal-tract shape parameters exceeds the number of acoustic parameters, the SM model is inherently nonunique if only the recorded speech signal is available.

In that context, it is important to emphasise that the SM model embodies the nonuniqueness of a completely lossless vocal-tract, perhaps in its most fundamental form. Whilst other, more physiologically-relevant parameterisations of the vocal-tract may afford more direct constraints on individual articulators such as the tongue body, the jaw, and the lips, the SM model captures the essential, acoustically-relevant manifestations of nonuniqueness in the entire shape of a completely lossless vocal-tract. It is also interesting to note that the acoustic-theoretic principles on which the SM model is founded, are equally the basis of the more recent, Distinctive Regions and Modes (DRM) model of Mrayati et al. (1988). Admittedly, the DRM model has already been used more extensively than the SM model: for example, in /V-V/ and /V-C-V/ trajectory modelling (Carré and Mrayati, 1991; Carré et al., 1992), in automatic recognition of place of articulation of plosives (Soquet and Saerens, 1994), in automatic recognition of vowels and /V-V/ sequences (Candille and Meloni, 1995), in real-time articulatory speech synthesis (Hill et al., 1995), and in acoustic-to-articulatory mapping itself (Richards et al., 1995). However, as the DRM model is founded on the same
acoustic principles (of a completely lossless acoustic tube) which are embodied in the
SM model, it does share the nonuniqueness properties of the latter, albeit without some
of the distinct advantages of the SM model’s parameter-set.

Indeed, one of the advantages of the SM model, is its inherently smooth
representation of vocal-tract area-functions, which therefore allows (cf. Equation 5.1)
computation of the cross-sectional area at any position x along the length of the vocal-
tract. The degree of smoothness itself is determined by the number of formants used in
the inversion; thus, as conjectured by Mermelstein (1967), the low-order formants can
be used to obtain a bandlimited version of an area-function. In that vein, the efficiency
of the SM model is manifest in its minimal set of parameters, which describe the vocal-
tract shape in terms of an orthogonal set of basis functions (the cosine terms in
Equation 5.1).

However, one of the disadvantages of the SM model\(^2\), is that the estimated vocal-
tract shape depends on the assumed vocal-tract length (VTL); Mermelstein (1967) has
already illustrated the potential sensitivity of the shape-length dependence. Based on
Zue’s (1969) thesis, Paige and Zue (1970) then proposed to optimise VTL such that the
resulting shape is least eccentric compared with a uniform area-function. Within the
analytical framework of the SM model, they obtained a quadratic expression for the
error criterion with respect to vocal-tract length, thus firmly establishing the existence
of a unique, global minimum for any given area-function. Furthermore, they were able
to re-estimate the vocal-tract lengths of Fant’s (1960) six Russian vowels, to within the
tolerance prescribed by Fant’s own, X-ray measurements (± 5 %). This combination of
theoretical elegance and empirical support has encouraged further use of the underlying
principle, which has since been referred to as “minimal articulatory antagonism”
(Lindblom and Sundberg, 1971), “minimal articulatory difference” (Bonder, 1983b), or
“minimum of muscle work” (Sorokin, 1992).

To summarise our preceding discussion, it is apparent that the great strength of
the SM model lies in its functional representation of vocal-tract shapes in terms of
orthogonal parameters, one-half of which are acoustic-phonetically significant by virtue

\(^2\) Indeed, of any articulatory model which retains the vocal-tract length as an explicit parameter.
of their quasi-linear, one-to-one mapping with the formant frequencies. However, the absence of resonance information (in the acoustic speech signal) pertaining to the required, second set of vocal-tract boundary conditions, implies that only the antisymmetric shape components can be inferred in practice. We therefore direct our attention in the next section, to the linear prediction (LP) vocal-tract model, which does inherently resolve the nonuniqueness problem without recourse to further, articulatory constraints.

5.2.2 Inherent Uniqueness in LP-derived Area-Functions

Indeed, as briefly reviewed in Chapter 2, the LP model does theoretically guarantee uniqueness in acoustic-to-articulatory mapping. In view of the inherent nonuniqueness of the completely lossless vocal-tract model reviewed in the previous section, we are therefore compelled to ask how the LP model secures uniqueness in vocal-tract shapes. In search of an answer to that question, we first consider the uniqueness issue as dealt with by the model’s two main proponents, namely Atal and Wakita.

Atal (1970) is widely credited (e.g., by Markel and Gray, 1976) for being the first to derive an acoustic tube model of the vocal-tract directly from the speech waveform, and to show that the formant frequencies and bandwidths are sufficient to uniquely determine the area-function in terms of a finite number of equal-length sections. In Atal’s version of the LP vocal-tract model, the formant bandwidths are brought about by a single, frequency-independent source of loss at the lip-end of an otherwise lossless acoustic tube. Invoking electrical-circuit analysis theory, Atal then proved (see, for example, Strube, 1977) that a discrete, $M$-section area-function is uniquely obtained through knowledge of the first $M$, discrete-time samples of the autocorrelation function of the pressure developed across the resistive lip termination, in response to a unit volume-velocity impulse at the glottis.

Atal and Hanauer (1971, Appendix F) then showed that an $M^{th}$-order, all-pole transfer function with all its poles inside the unit circle in the $z$-plane, can always be interpreted as that of a lossless acoustic tube with $M$ equal-length sections, terminated by a unit acoustic resistance at the lips. Furthermore, they defined a procedure for determining the area-function of such an acoustic tube, using the covariance method of
LP analysis. However, as pointed out by Markel and Gray (1976, p.79), they could make no claim “as to the applicability of the method for estimating vocal tract area functions”, owing to the potential instability of the covariance method of analysis, which may cause poles to lie on or outside the unit circle in the z-plane.

On the other hand, Wakita (1973) showed that the inverse filter obtained by the autocorrelation method of LP analysis, whose stability is theoretically guaranteed (Markel and Gray, 1976, p.103), is an equivalent representation of a lossless acoustic tube model with a resistive termination at the glottal end, and a short-circuit at the lips. Furthermore, he demonstrated that plausible vocal-tract shapes can be obtained directly from the speech waveform, provided certain analysis conditions (such as appropriate pre-emphasis) are applied.

Wakita and Gray (1975) then shed some light on the uniqueness of area-functions estimated by that method. In particular, they derived an expression for the lip impedance function of the M-section acoustic-tube model, both for a lossless and a lossy glottal termination. First, consistently with Borg’s (1946) proof reviewed in the previous section, Wakita and Gray (1975, p.579) showed that for a lossless model, the numerator and the denominator polynomials of the lip impedance function are independent of each other, such that “both of them are needed for the determination of a unique tube shape”; indeed, the roots of the two polynomials thus obtained are merely the resonance frequencies which satisfy, respectively, the two different sets of boundary conditions. By contrast, upon reinstatement of the lossy, resistive glottal termination, the two polynomials were found to be inter-dependent; it was shown that either of those polynomials can be obtained from the other, simply by changing the sign of all of the reflection coefficients. Indeed, while for the completely lossless model the two sets of eigenvalues (the roots of the numerator and denominator polynomials, respectively) are independent and real-valued (i.e., they have frequencies only), for the LP model with a resistive glottal termination they are dependent and complex-valued (i.e., they have both frequencies and bandwidths).

Whilst the formant bandwidths are thus clearly implicated in both Atal’s and Wakita’s explanations of the uniqueness of LP-derived area-functions, we sadly remain ignorant of the vocal-tract shape-related manifestations of uniqueness. In particular,
despite the implications clearly foreshadowed by the SM model, the literature apparently offers no insights regarding the roles of the antisymmetric and the symmetric components of LP-derived vocal-tract shapes, and their possible relations with the formant frequencies and bandwidths — we only know that by specifying the first $M/2$ formants, a discrete or step-wise, $M$-section area-function is obtained uniquely.

In that regard, it is important to note that the shape of an LP-derived area-function (similarly to that of a completely lossless model, as reviewed in the previous section) depends on its length. The principle of minimal articulatory effort which was applied first by Paige and Zue (1970), was then adapted to the LP method of inversion by Wakita (1977), who proposed to determine the vocal-tract length by minimising the eccentricity of the discrete LP area-function with respect to a uniform tube. Indeed, Wakita was able to re-estimate the lengths of five of Fant’s (1960) area-functions of six Russian vowels, with errors of the order of $\pm 5\%$ using the first four formant frequencies and bandwidths. Furthermore, he obtained reasonable estimates of the vocal-tract length of vowels recorded by several, adult male and female speakers of American English.

However, owing to the fact that the effective half-sampling frequency is inversely proportional to the length of each vocal-tract section, there remains the analysis artefact of an upper limit on allowed vocal-tract length (Wakita, 1977):

$$L_{\text{max}} = \frac{cM}{4F_{\text{hi}}}, \quad (5.6)$$

where $F_{\text{hi}}$ is the frequency of the highest formant considered, $M$ is the number of sections, and $c = 35300\text{ cm/sec}$ is the velocity of sound propagation in the vocal-tract. If for a given set of formant frequencies and bandwidths, there cannot be found an optimum vocal-tract length less than $L_{\text{max}}$, Wakita offers no alternative but to accept that upper limit (and the vocal-tract shape thus obtained).

Another limitation which is common to both the completely lossless and the LP-based inversion method, concerns the absolute value of estimated vocal-tract areas. For a completely lossless acoustic tube, the formants are independent of the value of the area scaling factor $A_0$ (cf. Equation 5.1); similarly for the LP acoustic tube model, which is completely specified by its reflection coefficients. Theoretically, $A_0$ should not
be so large as to violate the assumption of acoustic plane-wave propagation in the tract, and it should not be so small as to imply frication at the place of constriction for a vowel; physiologically, constraints on tongue-body volume suggest that $A_0$ should not have an excessive, phonetic range of variation. In practice, the area scaling factor is usually determined such that either the area at the glottal end, or the average area, or the vocal-tract volume, remains fixed.

Whilst both the LP and the SM model are thus capable of yielding only normalised\(^3\), or relative vocal-tract area-functions, there appears to be no evidence in the literature to suggest that this limitation presents an obstacle in using such estimated shapes to gain physiological insights into acoustic speech phenomena. On the contrary, a very small number of studies have even attempted to gain such insights using the LP method of inversion — these include Crichton and Fallside’s (1974) experimental system for deaf speech training; Gath and Yair’s (1988) successful identification of lingual “tremor” in the sustained sound /l/ recorded by Parkinsonian subjects; and Hansen and Womack’s (1996) plausible descriptions of the articulatory differences in the vowel /e/, in the word “help” recorded in neutral and angry states of emotional stress. Although the estimated shapes cannot be claimed to be exactly those produced by the speakers, the plausibility of the insights gained in such studies does suggest that the consequences of the LP model’s well-known limitations (as described, e.g., by Wakita, 1979; Sondhi, 1979) may not be as severe as the literature generally portrays.

In that context, perhaps the most misleading impression conveyed in the literature on LP-based inversion, is that area-functions can be estimated directly by LP analysis of the speech waveform, after applying appropriate pre-emphasis. Indeed, Wakita (1973, p.422) does suggest that even a simple pre-emphasis of 6 dB/octave is “essential”, in order to approximately equalise the combined spectral-slope effects of the glottal source and lip radiation (Fant, 1960), and thereby reduce the possibility of obtaining “unusual tract shapes”. More elaborate methods of counteracting the influences of source and radiation characteristics by flattening the spectral slope and enhancing the

\(^3\) Throughout this chapter, “normalised area-function” implies that the scale on the ordinate is dimensionless, owing to the fact that the area scaling factor ($A_0$ in Equation 5.1) is acoustically inconsequential; this should not be confused with speaker normalisation of area-functions, to which we refer in Chapter 6.
formant peaks, include the adaptive inverse filtering approach developed by Nakajima et al. (1973), the adaptive enhancing filter proposed by Tanaka and Nakajima (1975), and Fuchi’s (1977) use of the negative derivative of the LP phase spectrum (NDPS) to obtain an idealised, “stop/pass bands” spectrum. According to Sondhi (1979), “the only tenable conclusion is that the area recovery is very strongly dependent on the assumed source and radiation characteristics.”

Apart from the fact that the influences of source and radiation characteristics on the formants can sometimes be appreciable, it is rarely acknowledged that LP analysis usually yields a set of poles, amongst which may be found not only the formants, but also the so-called spurious poles. Whilst the LP spectrum may not be adversely affected by the presence of spurious poles (especially those of wide bandwidth), it is fair to assume that the shape of the LP area-function itself might be significantly affected — as far as the LP vocal-tract model is concerned, a spurious pole which lies in between two “true” formants is just another resonance of the acoustic tube. Implicit in Wakita’s (and Atal’s) descriptions of the LP-based method of inversion, is therefore the retention of only those poles which can be considered as the “true” formants.

However, the requirement of formants (estimation of which is itself non-trivial) does not resolve the issue of the relevance of measured formants to the LP model. Indeed, one of the major criticisms of the LP model is that it “lumps” all the sources of loss into a single, resistive termination (at the glottal end in Wakita’s model). By contrast, the formants measured from the acoustic speech waveform presumably include the effects of all losses which naturally occur in the human vocal-tract. For example (Fant, 1960; Flanagan, 1972), the glottal inductance tends to raise the centre-frequencies of formants which are more strongly affiliated with the pharyngeal cavity; the viscous and heat-conduction losses which are manifest along the surface of the vocal-tract walls, contribute mainly to the bandwidths of the higher formants; vocal-tract wall-vibrations tend to increase both the bandwidths and centre-frequencies of mainly the lower formants; and the lip-radiation impedance tends to increase the bandwidths of mainly the higher formants, and also to lower the centre-frequencies of formants which are more strongly affiliated with the oral cavity.

Wakita (1979) proposed to compensate for the LP model’s lack of sufficiently
realistic losses, by way of a “formant frequency conversion chart” based on formants synthesised with a more “realistic” (i.e., a more lossy) vocal-tract model. Similarly, Hafer and Coker (1975) had earlier attempted to correct measured formant frequencies prior to inversion, in order “to account for inductive yielding walls of the vocal tract”. However, it remains questionable to what extent such model-based formant correction procedures are applicable to real, measured data; and how viable it is to formulate a similar correction procedure for formant bandwidths (which are, after all, more significantly affected than the formant frequencies by the LP model’s simplified assumptions in regard to vocal-tract losses).

Despite these limitations, the LP-based method of inversion is at an advantage compared with many other such methods, owing to the fact that the LP model is an acoustic analysis model, whereby conversion of formants into an area-function is direct (or non-iterative) and computationally inexpensive. However, as suggested by Broad and Shoup (1975, quoted in Section 2.4.3.2), the potential of the LP vocal-tract model to yield physiological insights into acoustic speech phenomena has to date been shamefully under-exploited. Perhaps a determining factor in the overwhelming neglect of the LP model over the past two decades, is its very coarse, step-wise representation of area-functions, from which it is admittedly difficult to extract crucial articulatory landmarks such as the place of constriction. More importantly, the coarse representation inhibits quantitative comparisons of area-functions of potentially different lengths, whether of different vowels or different speakers.

Those types of problems might conceivably be resolved by smoothing the discrete LP area-function, and thereby obtaining a continuous and parameterised representation. Indeed, Wakita and Gray (1975, Figure 4) used a Chebyshev polynomial approximation (also used earlier by Nakajima et al., 1973) to illustrate the LP-derived area-functions of five American English vowels. Although Wakita and Gray (1975, p.578) thus obtained “visually satisfying results” with the places of constriction more precisely defined, their particular choice of smoothing function was admittedly “somewhat arbitrary”.

On the other hand, as reviewed in the previous section, the SM model does afford a smooth representation of vocal-tract shapes, and more importantly, a resonance-based parameterisation. Whilst it may therefore seem advantageous to use the SM model to
parameterise LP-derived area-functions, it is not immediately obvious whether the acoustically-relevant components of LP-derived vocal-tract shapes are equivalent to those of a completely lossless vocal-tract on which the SM model is based. In the next section we seek to answer that question, by attempting to identify the parameters of unique, LP-derived area-functions.

5.3 Parameters of Unique LP-derived Area-Functions

In Section 5.1 we defined the two main criteria which our method of inversion would need to satisfy — namely, uniqueness and formant-based parameterisation. Our bipartite rationale then addressed these issues by invoking two classic, but sadly over-neglected models, which do promise to fulfil our requirements. First, we reviewed the SM model, which does indeed provide an acoustically-meaningful parameterisation of the vocal-tract area-function. However, its underlying assumption of a completely lossless vocal-tract, implies that a unique area-function can only be obtained if both the zeros (formants) and the poles of the lip impedance function are known. This fundamental limitation then led us to consider the LP vocal-tract model, which is the only one to guarantee uniqueness of area-functions estimated using acoustic parameters, all of which can be measured from the recorded speech signal. However, whilst both the formant frequencies and the bandwidths are implicated in the LP model’s uniqueness, it has never been shown how LP-derived vocal-tract shapes exploit the information contained in the formants, and of how those shapes might best be parameterised in an acoustically meaningful way.

As far as we are aware, Wakita and Gray (1975) were the first and only researchers to attempt to relate the LP model with the earlier work of Schroeder (1967) and Mermelstein (1967). Indeed, they suggested that the LP and the SM models could be used interchangeably to obtain a unique vocal-tract shape, once the appropriate acoustic parameters had been identified by LP analysis. In particular, they proposed that the frequencies of the zeros and poles of the lip impedance function (as required by the SM inversion method) could be found after LP analysis of the speech waveform, by setting the glottal reflection coefficient to unity in order to render the LP vocal-tract model completely lossless, then finding the required zeros of the lip impedance function.
by solving for the roots of the modified LP polynomial, and similarly the required poles, after changing the sign of all of the reflection coefficients. It is interesting to note that the LP line spectrum pair (LSP) parameters (Itakura, 1975), which are often regarded merely as an alternative set of LP acoustic parameters, are indeed the inter-leaving frequencies of the zeros and poles of the lip impedance function of a completely lossless LP vocal-tract model, and are therefore precisely the necessary acoustic parameters which Wakita and Gray (1975) suggested for obtaining a unique vocal-tract shape using the SM inversion method.

Does this imply that the poles of the lip impedance function are contained in the acoustic speech signal after all, and that the SM model can therefore be used to obtain unique vocal-tract shapes without resort to the lip impedance-tube method proposed by Schroeder (1967)? To answer that question, it is important to note that Wakita and Gray’s (1975) proposed method of uniting the LP and the SM models, relies on post-analysis modification of the LP glottal reflection coefficient; subsequent conversion to LSP frequencies does not change the shape of the LP area-function, which itself is uniquely determined by the frequencies and bandwidths of the LP poles yielded in the original analysis. The poles of the lip impedance function thus obtained, are therefore only indirectly inferred from the LP area-function, rather than directly measured from the acoustic speech signal.

It emerges from the preceding discussion, that our knowledge of the uniqueness properties of LP-derived vocal-tract shapes is still rather limited. Although in their respective formulations of the LP-based inversion method, Atal and Wakita certainly provide evidence of the uniqueness of LP-derived area-functions, they stop short of describing how the components of estimated vocal-tract shapes themselves are involved in the uniqueness. By contrast, the SM model lends direct insights into those fundamental components of vocal-tract shapes which are of greatest relevance from acoustic-phonetic, perceptual, and articulatory points of view. Notwithstanding the differences between the respective vocal-tract models, it would therefore seem appropriate to use the SM model as a basis for identifying the parameters of unique LP-derived area-functions.

As reviewed in the previous section, the underlying difference between the SM
model and the LP acoustic-tube model is that the latter includes a resistive element at the glottal end. The acoustic consequences of a purely resistive glottal termination have been studied both theoretically (e.g., Fant, 1960; Flanagan, 1972) and empirically (e.g., Badin and Fant, 1984). As a result, it is well-established that a resistive glottal termination influences the formant frequencies far less than the bandwidths; furthermore, that the resulting losses are frequency-independent. We might therefore expect that the LP model shares the distinctive articulatory-acoustic relations embodied in the SM model — in particular, that it inherits the property of a quasi-linear relation between formant frequencies and the corresponding, antisymmetric shape components of the logarithmic area-function. We address this issue in Section 5.3.1, by presenting new results which shed light on the formant frequency-dependence of LP-derived vocal-tract shapes. In Section 5.3.2 we then address the question of fundamental concern, whether the formant bandwidths, which provide the crucial second-half of acoustic information required to secure uniqueness of LP-derived vocal-tract shapes, are at all related to the undetermined, symmetric components of those shapes.

### 5.3.1 Dependence of LP-derived VT-Shapes on Formant Frequencies

Our first step towards marrying the SM model and the LP model, is to prove their equivalence in regard to the dependence of vocal-tract shapes on the formant frequencies. This proof is first considered from a theoretical point of view (in Section 5.3.1.1), with the simplifying assumption of a two-section vocal-tract area-function. It is then extended (in Section 5.3.1.2) with empirical results which substantiate our hypothesis of the equivalence of the two models, and provide the most extensive validation of the SM model ever to appear.

#### 5.3.1.1 Partial Theoretical Proof

One of the distinguishing features of the completely lossless SM model reviewed in Section 5.2.1, is its continuous representation of vocal-tract area functions, in terms of the variable $x$. Hence, each parameter of the SM model specifies a component of the vocal-tract shape, which is independent of the number of sections used to implement the model in practice (as would be required, for example, to synthesise the acoustic
resonances of a given area-function using numerical procedures). The LP model, on the other hand, not only represents the area-function in terms of a discrete number of lossless sections $M$ (which, in the absence of real-valued poles, is equal to twice the number of LP poles effectively used), but is also characterised by a resistive (i.e., a lossy) termination at one end of the vocal-tract. In order to confirm our hypothesis that LP-derived vocal-tract shapes are governed by the same basic principles that define the SM model, it is therefore necessary first to recast the main result of the SM model in terms of a discrete-sectioned, lossless acoustic tube.

Bonder (1983a) has shown that Webster’s Horn Equation (Equation B.3 or B.4, in Appendix B) can be used to derive an analytical solution for the resonances of a lossless acoustic tube with up to 10 equal-length sections. For simplicity, we will consider a single-resonance acoustic tube which, according to the LP model, comprises only two sections of equal length $L/2$. The simplified version of Bonder’s (1983c) so-called “$n$-tube formula” then reduces to the following expression (cf. also Fant, 1960, p.65; Flanagan, 1972, p.70):

$$\tan^2 \left( \frac{\pi LF_1}{c} \right) = \frac{A_1}{A_2},$$  \hspace{1cm} (5.7)

where $A_1$ and $A_2$ are the cross-sectional areas of the front and the back vocal-tract sections, respectively, and $F_1$ is the frequency of the single resonance.

According to the SM model (Equation 5.2), the most efficient way to perturb $F_1$ from its neutral value is to perturb the uniform area-function according to the antisymmetric shape parameter $a_1$. For a two-section acoustic tube, Equation 5.1 then yields the following area-perturbations which would induce a positive perturbation in the first formant frequency:

$$A_m = e^{-a_1 \cos((2m-1)n/4)}, \hspace{1cm} m = 1, 2 \hspace{1cm} (5.8)$$

where it is assumed that the area of each section takes on the value given by Equation 5.1 at the centre of that section (i.e., at $x = 3L/4$ and $x = L/4$ for the front and the back sections, respectively), and that $A_0 = 1$. Equations 5.8 and 5.7 then together yield the following expression for the single resonance-frequency $F_1$ in terms of the vocal-tract shape parameter $a_1$: 

204
\[ F_1 = \frac{c}{\pi L} \tan^{-1}\sqrt{e^{-na\sqrt{2}}} . \]  

(5.9)

This last result, which is the homologue of Equation 5.2 for \( n = 1 \), can be regarded as an \textit{exact} version of the SM model for a lossless two-tube. It is exact in that it correctly predicts the value of the single resonance frequency as it would be obtained by a sufficiently accurate numerical procedure to solve Webster’s Horn Equation for a two-section area-function, independently of the degree of eccentricity of that area function with respect to a uniform tube. Assuming \( c = 35300 \, \text{cm/sec} \) and \( L = 17.65 \, \text{cm} \) as in our earlier illustration of the SM model (Figure 5.1), the solid curve in Figure 5.2 illustrates the result just derived, showing the predicted variations in \( F_1 \) about its neutral value of 500 Hz, for area-function perturbations over a fairly wide range \( a_1 \in [-2, +2] \). The exactness of Equation 5.9 is indeed evidenced by the way in which it correctly predicts the \( F_1 \) asymptotes of 0Hz and 1000Hz, respectively, for large positive and large negative perturbations of \( a_1 \). By contrast, the linearity of the SM model (Equation 5.2 for \( n = 1 \), shown here by the dotted curve) is strictly only valid for relatively small perturbations about a uniform area-function.

Having established the exact form of the SM model for a completely lossless area-function with two sections, we now proceed to derive an analogous expression for the LP model. According to Wakita & Gray’s (1975) formulation, the single resonance of a two-tube LP-model is given by the complex-conjugate roots of the following numerator-polynomial of the lip impedance function:

\[ z^2 + \mu_1(1 + \mu_2)z + \mu_2 = 0, \]

(5.10)

where \( \mu_1 = (A_1 - A_2) / (A_1 + A_2) \) is the reflection coefficient which describes the shape of the two-section area-function, \( \mu_2 = \rho c / A_e^{LP} \) is the glottal reflection coefficient which determines the effective (LP) glottal opening area \( A_e^{LP} \) and hence the losses introduced by the glottal resistance, and \( z = e^{2\pi LF/c} \). Solving for the roots of the quadratic Equation 5.10, the single resonance frequency is then expressed in terms of the two reflection coefficients, as follows:

\[ F_1 = \frac{c}{2\pi L} \tan^{-1}\left( \frac{\sqrt{4\mu_2 - \mu_1^2(1 + \mu_2)^2}}{-\mu_1(1 + \mu_2)} \right), \]

(5.11)
where the range of the inverse-tangent function is taken as \([0, \pi]\) by adding \(\pi\) to the result whenever the argument is negative-valued, in order to ensure that only the pole lying in the top half of the \(z\)-plane is considered.

Finally, to obtain an expression analogous to Equation 5.9, we need to derive the relation between the reflection coefficient \(\mu_1\) and the area-function parameter \(a_1\). Making use of Equation 5.8 and the definition of the reflection coefficient in terms of the individual areas, we obtain the following:

\[
\mu_1 = -\tanh\left(\frac{a_1}{\sqrt{2}}\right). \tag{5.12}
\]

Substitution of Equation 5.12 into Equation 5.11 then yields the desired relation which is analogous to the earlier-derived Equation 5.9, with the extra flexibility afforded by the resistive loss-element at the glottal end of the two-tube LP-model — an infinite glottal impedance (i.e. a closed glottis) can be obtained by setting \(\mu_2 = 1\), while finite and increasing losses are introduced as \(\mu_2\) decreases towards zero.

Figure 5.2: Theoretical behaviour of the resonance frequency \(F_1\) of a 2-tube, as a function of the shape-perturbation parameter \(a_1\). Solid curve: exact relation (Equation 5.9) derived from a completely lossless 2-tube. Dashed curves: exact relation (Equations 5.11 and 5.12) derived from a 2-section LP model, shown here for three values of the glottal reflection coefficient \(\mu_2=\{1.0, 0.7, 0.4\}\). Dotted curve: linear relation predicted by the SM model (Equation 5.2).
The dashed curves in Figure 5.2 illustrate the LP-based equations just derived, for three different values of the glottal reflection coefficient $\mu_2 = \{1.0, 0.7, 0.4\}$. As expected, when $\mu_2 = 1$ (completely lossless, two-tube LP-model) the dashed curve coincides exactly with the solid curve, which was obtained earlier from the exact version of the two-tube SM model (Equation 5.9). The dashed curve for $\mu_2 = 0.7$ shows a very similar behaviour of the resonance frequency, indicating that the inclusion of a moderate amount of loss at the glottal end does not significantly alter the relation between $F_1$ and the shape of a two-tube. However, it is interesting to note that as the glottal resistance is made to play a more significant role (dashed curve for $\mu_2 = 0.4$), the behaviour of $F_1$ in the two-tube LP-model begins to approach the quasi-linear behaviour (dotted curve) predicted by the original SM model (Equation 5.2), which is based on a completely lossless acoustic-tube, and which itself makes no assumptions as to the number of vocal-tract sections.

We should emphasise, however, that the range over which $a_1$ has been perturbed in Figure 5.2, is much wider than would normally be expected for vocalic configurations (for example, Mermelstein’s (1967) analysis of Fant’s (1960) six Russian vowels yielded values for $a_1$ spanning approximately $[-1, 1]$, i.e., about half of the range shown here). It is also quite clear in Figure 5.2 that the small but systematic differences between the theoretically-derived models, are only of significance for such large perturbations as are shown at the positive and the negative extremes along the abscissa. Within the more restricted range of perturbations (e.g., $[-1, 1]$) which are typical of vocal-tract shapes for vowels, Figure 5.2 therefore confirms that the behaviour of the resonance frequency $F_1$ as a function of the shape-perturbation parameter $a_1$ is essentially common to both models. Our results concerning the two-section vocal-tract, thereby support a partial theoretical proof of the validity of the SM model within the LP modelling framework.

5.3.1.2 Empirical Validation

We now aim to further substantiate the partial theoretical proof presented thus far, by extending our validation to vocal-tract shapes with more than two sections. Two methods are used to illustrate empirically the relation between formant frequencies and
antisymmetric (and symmetric) perturbations of an LP area-function. In particular, we show that this relationship is essentially the same as that predicted by the SM model for a completely lossless acoustic tube.

The first method is similar to our earlier illustration of the SM model (Figure 5.1, in Section 5.2.1), and involves nomograms depicting formant frequency variations as a function of perturbations in a single SM-model parameter at a time. However, in contrast to the earlier results which were generated using a completely lossless vocal-tract model, the resonances of computed vocal-tract shapes are here synthesised using the LP vocal-tract model, which not only involves a discrete number of equal-length sections (strictly equal to twice the number of formants), but also includes a resistive termination at the glottal end. Hence, for each perturbation of an SM-model parameter $a_n$, an area-function with $M$ equal-length sections is first computed, with the area of each section taking on the value given by Equation 5.1 at the centre of that section, and assuming $A_0 = 1$. The discrete and normalised area-function $A_m$, $m = 1, \ldots, M$ is then converted to a set of reflection coefficients $\mu_m$, $m = 1, \ldots, M-1$, and the glottal (the $M$th) reflection coefficient is chosen such that $0 < \mu_M \leq 1$. A well-known recursive algorithm (Markel and Gray, 1976) is then used to transform the set of $M$ reflection coefficients to a set of $(M+1)$ LP autoregressive coefficients, which defines the $M$th-order polynomial of the LP inverse filter. Finally, the $M/2$ formant frequencies are obtained by solving for the roots of the LP inverse-filter polynomial.

Figure 5.3 shows a set of nomograms obtained for LP area-functions of $M = 8$ sections, resulting in the synthesis of only the first four formants. As in Figure 5.1, the panels on the left-hand side depict formant-frequency nomograms for perturbations of each of the first three odd-indexed parameters ($a_1$, $a_3$, and $a_5$) separately, while those on the right-hand side show perturbations of each of the first three even-indexed parameters ($a_2$, $a_4$, and $a_6$). Also shown in each panel are the sets of superimposed, eight-section LP area-functions, each with a fixed total length $L = 17.65 \text{ cm}$, whose shapes are determined according to the single parameter being perturbed about its neutral (zero) value. Rather than the usual, step-wise depiction of LP area-functions,

---

4 This well-known sequence of procedures for obtaining the formants of a given, discrete area-function, will henceforth be referred to as “LP synthesis”.
Figure 5.3: Nomograms depicting changes in the first four formant frequencies, as a function of
perturbations in each of the first six, vocal-tract shape-parameters of the SM model. Each M=8-section
area-function is of length $L=17.65\text{cm}$, and the area scaling factor is $A_0=1.0$; the formants are
synthesised using the LP vocal-tract acoustic model, with the glottal reflection coefficient $\mu_M=1.0$
(solid curves), 0.7 (dashed curves), and 0.4 (dotted curves).
we emphasise their shapes by plotting a piecewise-linear representation obtained by joining the section-centres; these area-functions are therefore merely a more coarsely-sampled version of those shown earlier in Figure 5.1.

For each formant in Figure 5.3, are shown three curves which have been obtained, respectively, using a completely lossless implementation of the LP vocal-tract model ($\mu_M = 1$, shown by the solid curve), and two other configurations with increasing amounts of loss ($\mu_M = 0.7$, shown by the dashed curve; and $\mu_M = 0.4$, shown by the dotted curve). It is evident that in most cases, the dashed and the solid curves are indistinguishable — only in a few cases, mainly in the higher formants and for extreme values of perturbation, are the dotted curves visually separable from those pertaining to a more tightly-closed glottis. Figure 5.3 thus confirms that the resistive glottal termination has only a very small influence on the LP-synthesised formant frequencies. In particular, as the equivalent glottal opening area is increased by decreasing the value of $\mu_M$, a slightly steeper (more negative) slope is obtained for the first-order relation between each $F_n$ and the corresponding shape parameter $a_{2n-1}$. Nevertheless, the general behaviour of the formant-frequency nomograms is similar to that observed in Figure 5.1, with each of the odd-indexed Fourier cosine coefficients claiming first-order control of a unique formant frequency, and with the even-indexed parameters exerting only second-order influence by comparison. We therefore conclude that these results are in agreement with the theoretical predictions of formant frequency behaviour based on the SM model (Equation 5.2).

In our second, empirical validation of the equivalence of the SM and LP models, we aim to reveal the strength and the inherent structure of the relations between shape and resonance parameters, by considering a more extensive range of shape-parameter perturbations about a neutral tube. Retaining the eight parameters ($a_n$, $n = 1, \ldots, 8$) which are implied by the SM model (Equation 5.1) for an 8-section area-function, a total of 6561 different vocal-tract shapes are generated by all possible combinations of ternary perturbations of each parameter ($-0.1, 0.0, +0.1$). The area-functions are found using Equation 5.1, and assuming a fixed vocal-tract length $L = 17.65$ cm and a normalised area-scaling factor $A_0 = 1$; a completely lossless acoustic tube is assumed ($\mu_s = 1$), and the first four formant frequencies are obtained using the LP forward
routines. Finally, the coefficient of correlation (or Pearson’s $r$) is computed across all 6561 configurations, between the shape parameters and the relative formant frequencies $F_{n}^{(rel)} = (F_{n} - F_{n}^{(neut)}) / F_{n}^{(neut)}$ (i.e., each formant frequency normalised with respect to its neutral value $F_{n}^{(neut)} = (2n-1)c / 4L$ for the given vocal-tract length, as implied in Equation 5.2).

Table 5.1 shows the correlations thus obtained. As expected, the table is heavily dominated by the entries along the main diagonal. Whilst the SM model theoretically establishes the negatively-sloped and quasi-linear relation between the $n^{th}$ relative formant frequency and the corresponding shape parameter $a_{2n-1}$, the high, negative correlations along the main diagonal of Table 5.1 summarise the strength of that linearity. In addition, the relatively insignificant correlations in all of the off-diagonal entries of Table 5.1, support the notion of a predominantly one-to-one mapping between each odd-indexed Fourier cosine parameter and the corresponding formant frequency. As far as we are aware, this evidence provides the most extensive, empirical validation of the SM model ever reported. More importantly, our results further validate the equivalence of the SM and the lossless LP model$^5$.

Indeed, the theoretical and empirical evidence presented in this section supports our hypothesis that the LP model and the SM model are isomorphic in regard to the fundamental relation between formant frequencies and the antisymmetric components of the logarithmic area-function. This mutual congruence in the basic acoustic properties

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_3$</th>
<th>$a_5$</th>
<th>$a_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1^{(rel)}$</td>
<td>-0.996</td>
<td>-0.001</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>$F_2^{(rel)}$</td>
<td>-0.001</td>
<td>0.991</td>
<td>-0.004</td>
<td>-0.001</td>
</tr>
<tr>
<td>$F_3^{(rel)}$</td>
<td>0.000</td>
<td>-0.004</td>
<td>0.986</td>
<td>-0.010</td>
</tr>
<tr>
<td>$F_4^{(rel)}$</td>
<td>0.000</td>
<td>-0.001</td>
<td>-0.007</td>
<td>0.997</td>
</tr>
</tbody>
</table>

Table 5.1: Linear correlation coefficients (Pearson’s $r$) between each of the first four (relative) formant frequencies LP-synthesised from 6561 perturbations of an 8-section, uniform area-function, and each of the first four odd-indexed VT-shape parameters from which those area-functions were generated.

$^5$ A complementary proof of the equivalence of those two models, is to show that the LP method of inversion can be “stepped down” such as to yield area-functions which would have been obtained using the SM model, starting from the same set of formant frequencies. The interested reader is referred to Appendix D.
of the two models can therefore be used to advantage, by adopting a parameterisation of LP-derived area-functions in terms of the Fourier cosine coefficients of the SM model (cf. Equation 5.1). Insofar as this type of parameterisation allows a smooth and continuous representation of vocal-tract shapes, it immediately resolves the difficulties of accurately identifying crucial articulatory landmarks such as the place of constriction, and suggests ways to approach the problem of quantitatively comparing two or more LP-derived area-functions (elaborated in Section 5.5.1). However, the key strength of the SM-model parameterisation lies in its acoustically-meaningful description of vocal-tract shapes, which has hitherto never been attempted within the LP modelling framework.

5.3.2 Dependence of LP-derived VT-Shapes on Formant Bandwidths

Thus far, we have been fortunate to be able to recall the SM model and unfold it as far as elucidating the acoustic determinants of the antisymmetric components of LP-derived vocal-tract shapes. However, despite the acknowledged importance of the formant bandwidths in securing uniqueness in the LP-based method of inversion, it has never been explained how these acoustic parameters contribute to the shape of LP-derived area-functions. In the next two sections, we therefore consider this problem from a theoretical and an empirical point of view, respectively.

5.3.2.1 Theoretical Motivation

In this section we present a theoretical derivation of the relation between the formant bandwidths and the glottal reflection coefficient of the LP vocal-tract model. Whilst a similar derivation was presented briefly by Kasuya and Wakita (1979), we herein present a more detailed derivation which leads to an expression involving the mean bandwidth, and which then allows us to foreshadow the bandwidth-dependence of LP-derived vocal-tract shapes.

According to the well-known recursive algorithms for converting between the polynomial (autoregressive) coefficients and the reflection coefficients (e.g., Markel and Gray, 1976), the glottal reflection coefficient $\mu_M$ is equal to the $M^{th}$ coefficient of the inverse-filter polynomial (see for example Equation 5.10 which shows that, for a
second-order polynomial in $z$, the constant term is equal to the highest-indexed reflection coefficient $\mu_2$. In the same vein, it is relevant to recall Wakita’s (1973) statement that, as the LP inverse-filter polynomial has only real or complex-conjugate roots, the coefficient of its highest-order term ($z^{-M}$) is determined only by the bandwidths of those roots. Therefore it is not surprising that the bandwidths should determine the value of the glottal reflection coefficient, which itself defines the effective glottal opening area $A_{M+1}$, and hence the amount of losses dissipated through the characteristic resistance of the glottal termination.

Indeed, a well-known theorem of algebra (e.g., Chrystal, 1964, p.432) states that, for a monic polynomial (such as that of the LP inverse-filter), the constant term (or the $M^{th}$ autoregressive coefficient, which is equal to the glottal reflection coefficient $\mu_M$) is proportional to the product of all $M$ roots of the polynomial, as follows:

$$\mu_M = (-1)^M \prod_{m=1}^{M} z_m. \quad (5.13)$$

Without loss of generality, the LP order of analysis $M$ can be assumed to be even, thereby reducing the constant of proportionality to positive unity. Further, since the roots of the LP polynomial occur in complex-conjugate pairs, the product of all of the complex roots is equal to the product of their magnitudes, such that Equation 5.13 can be rewritten as follows:

$$\mu_M = \prod_{m=1}^{M} |z_M|. \quad (5.14)$$

It is also well known that the bandwidth of the $m^{th}$ root of the LP inverse-filter polynomial is proportional to the natural-logarithm of the magnitude of that root in the $z$-plane, and is given by the following:

$$B_m = \frac{F_s}{\pi} \ln |z_m|, \quad (5.15)$$

where $F_s$ is the sampling frequency. The sum of the bandwidths of all $M$ roots can therefore be expressed as follows:

$$\sum_{m=1}^{M} B_m = \frac{F_s}{\pi} \ln \left( \prod_{m=1}^{M} |z_m| \right). \quad (5.16)$$
Substituting for the product of the magnitude of the roots from Equation 5.14, and noting also that $F_c = cM / 2L$ (which is a property common to all vocal-tract models with a discrete number of sections, such as the LP model), Equation 5.16 can finally be rewritten as follows:

$$\bar{B} = -\frac{c}{2\pi L} \ln(\mu_M),$$

(5.17)

where $\bar{B}$ denotes the mean resonance bandwidth. According to this last result, the mean bandwidth of the $M/2$ independent poles of an $M$:th-order LP analysis, is proportional to the logarithm of the glottal reflection coefficient, and is inversely proportional to the assumed vocal-tract length.

More importantly, Equation 5.17 states that, for a fixed vocal-tract length, the glottal reflection coefficient is the sole determinant of the *mean* bandwidth of the formants generated by an LP area-function. This implicates the remaining reflection coefficients ($\mu_1, \ldots, \mu_{M-1}$) which define the shape of the vocal-tract, in determining the value of the individual formant bandwidths relative to the mean bandwidth $\bar{B}$ for the given LP area-function. In the next section we demonstrate the truth of this implication, and as a result, empirically derive a new set of vocal-tract shape parameters which relate distinctively to the *relative formant bandwidths*, analogously to the SM model (Equation 5.2) which relates each odd-indexed Fourier cosine coefficient to a unique, *relative formant frequency*.

### 5.3.2.2 Empirical Justifications

With this clear objective in mind, we now endeavour to explore the dependence of LP-derived vocal-tract shapes on the formant bandwidths. In particular, we embrace the theoretically-motivated framework of the SM model, and seek to gain insights into the bandwidth-dependence of shapes perturbed about a neutral tube. However, rather than venture a hypothesis on the type of perturbation which might be necessary to induce a distinctive change in only a single formant bandwidth, we use the LP inverse method to obtain the vocal-tract shapes corresponding to such acoustic perturbations.

In order to obtain vocal-tract shapes perturbed from the neutral configuration according to variations in a single bandwidth at a time, we first need to identify the
formant frequencies and bandwidths of a uniform LP area-function. A cursory examination of the LP-based nomograms presented earlier (in Figure 5.3) reveals that the neutral-tube formant frequencies of the LP model are basically the same as those of a completely lossless uniform tube of the same length $L$, as given by the quarter-wavelength formula; i.e., $F_n = (2n-1)c/4L$, regardless of the value of the glottal reflection coefficient. As the resistive glottal termination of the LP acoustic-tube model induces energy losses which are frequency-independent, the values of the $M/2$ bandwidths of a uniform tube represented by $M$ equal-length sections are identical to the mean bandwidth, for which we have derived an exact relation (Equation 5.17). For example, an $M$-section, uniform LP area-function of total length $L = 17.65$ cm and glottal reflection coefficient $\mu_M = 0.7$, has the following set of formant parameters: $F_n = (2n-1)500$ Hz and $B_n \approx 113.5$ Hz, for $n = 1,\ldots,M/2$.

The bandwidth-dependence of LP-derived vocal-tract shapes can now be assessed by perturbing a selected bandwidth, while keeping all other formant parameters (i.e., the frequencies and the remaining bandwidths) fixed at their respective, neutral values. Assuming a nominal, neutral bandwidth of 100 Hz, each formant bandwidth is chosen in turn, and perturbed two steps below (50 Hz and 80 Hz) and two steps above (125 Hz and 200 Hz) the neutral value. The LP inverse method is then used to obtain a vocal-tract shape for each set of formant parameters, by the well-known transformations through LP autoregressive coefficients and LP reflection coefficients. As our aim is to discern an underlying pattern in the vocal-tract shapes obtained in this way, the first seven formants are specified in order to secure $M = 14$ vocal-tract sections, which does permit a satisfactory spatial resolution for visually identifying the general form of the area-function perturbation.

Figure 5.4 shows the 14-section, LP-derived vocal-tract shapes obtained by the procedure just described, for changes in each of the first three formant bandwidths. It is immediately apparent that perturbation of only the $n^{th}$ formant bandwidth, induces an approximately sinusoidal perturbation of the area-function, with $2n-1$ half-cycles within the length of the vocal-tract from the glottis to the lips. Recall from Section 5.2.2 that the area scaling factor $A_0$ is acoustically inconsequential in both the completely lossless and the LP vocal-tract model; hence, the area-functions are plotted in such a
Figure 5.4: Bandwidth-dependence of LP-derived vocal-tract shapes. Each formant bandwidth $B_n$ is perturbed over five values {50, 80, 100, 125, 200} Hz, with all other bandwidths fixed at the nominal, neutral value 100 Hz; the formant frequencies are fixed at neutral values $F_n = (2n-1)c/4L$, and the vocal-tract length is fixed at $L=17.65$ cm. The top, middle, and bottom graphs show the LP-derived area-functions for perturbations in $B_1$, $B_2$ and $B_3$, respectively. Seven formants are used in order to obtain 14-section LP area-functions (solid lines); at about the mid-length of each area-function, the largest through smallest area corresponds to the increase in bandwidth from 50 Hz to 200 Hz as listed above. Dashed curves: envelopes of the 1st, 3rd, and 5th sinusoidal components for the given $L$, superimposed on the step-wise area-functions in order to emphasise the functional form of the shape perturbations.
way as to visually emphasise the sinusoidal component, by setting the mean logarithmic area to zero in the bottom two graphs, and by setting a constant area for the first section in the top graph.

The significance of the LP-based result portrayed in Figure 5.4, is perhaps best appreciated in the context of the SM model. In Section 5.3.1 we established a link between the SM model and the LP model, by showing that the latter does inherit the fundamental relation between antisymmetric shape components and perturbations in the formant frequencies. In Figures 5.1 and 5.3 we then showed that for both the SM and LP models, respectively, the formant frequencies are relatively independent of the symmetric shape components. Remarkably, Figure 5.4 reveals that when a completely lossless vocal-tract 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 resulting formant bandwidths. Our results therefore elucidate the bandwidth-dependence of LP-derived area-functions, and furthermore suggest a functional form of the required, symmetric shape perturbations (as shown in Figure 5.4 by the superimposed, dashed curves), in terms of the odd-indexed coefficients of the Fourier sine (rather than the SM model's original Fourier cosine) series of the logarithmic area-function.

Having thus identified the shape components which fundamentally relate to the individual bandwidths, a more complete assessment of the implied, shape-resonance relations is afforded by nomograms prepared using the LP model. Considering only the first four formants, LP area-functions are each represented by \( M = 8 \) sections of equal length. Each of the shape parameters (i.e., the odd-indexed Fourier cosine coefficients \( a_{2n-1} \) and the newly-implicated, odd-indexed Fourier sine coefficients, henceforth denoted as \( b_{2n-1} \)) is perturbed across the range \([-0.8, +0.8]\) in steps of 0.05, while keeping all other parameters fixed at zero value. Either discrete cosine expansion (for the \( a_{2n-1} \)) or discrete sine expansion (for the \( b_{2n-1} \)) is then used to obtain 8-section area-functions of fixed length \( L = 17.65 \text{ cm} \), from which the first four formant frequencies and bandwidths are LP-synthesised (assuming a nominal value for the glottal reflection coefficient \( \mu_M = 0.7 \) which, according to Equation 5.17, yields a reasonably realistic, mean bandwidth of \( \bar{B} \approx 113.5 \text{ Hz} \)).
Figure 5.5(a): Formant-frequency nomograms generated by perturbing a uniform, 8-section LP area-function, according to each of the first four, odd-indexed Fourier cosine ($\Delta a_{2n-1}$, left graphs) and sine ($\Delta b_{2n-1}$, right graphs) shape-components, for $n=1$ (diamond symbols joined by green lines), $n=2$ (plus symbols joined by blue lines), $n=3$ (square symbols joined by red lines), and $n=4$ (cross symbols joined by black lines).
Figure 5.5(b): Formant-bandwidth nomograms generated by perturbing a uniform, 8-section LP area-function, according to each of the first four, odd-indexed Fourier cosine ($\Delta a_{2n-1}$, left graphs) and sine ($\Delta b_{2n-1}$, right graphs) shape-components, for $n=1$ (diamond symbols joined by green lines), $n=2$ (plus symbols joined by blue lines), $n=3$ (square symbols joined by red lines), and $n=4$ (cross symbols joined by black lines).
Figure 5.5(a) shows the influence of each of the first four parameters $a_{2n-1}$ (left graphs) and $b_{2n-1}$ (right graphs) on each of the first four formant frequencies. It is quite clear that the first-order influence of each parameter $a_{2n-1}$ is a negatively-sloped, quasi-linear variation in the corresponding formant frequency $F_n$. By comparison, the influence of each of the remaining Fourier-cosine shape parameters on that formant frequency, as indeed of each of the Fourier-sine shape parameters shown in the right-hand graphs, is relatively minor. The largest of these second-order effects appear in the $F_3$ nomograms for perturbations in $a_7$ (left graph) and in $b_7$ (right graph), which describe the antisymmetric and the symmetric shape-component, respectively, of highest spatial resolution for the 8-section LP model. An interesting phenomenon is observed in the bottom-left graph of Figure 5.5(a), where there appears a discontinuity in the nomogram for $F_4$, as $a_7$ is reduced below about $-0.5$. This is merely an artefact of the 8-section vocal-tract model, which prohibits measurement of formants higher than the effective upper spectral limit (or half-sampling frequency), equal to 4kHz for the given vocal-tract length. Notwithstanding such analysis artefacts, the nomograms in Figure 5.5(a) further support our earlier evidence concerning the distinctive influence of antisymmetric vocal-tract shape perturbations on the formant frequencies. Furthermore, they confirm that symmetric shape perturbations caused by variations in the odd-indexed Fourier sine components (which we earlier associated with the formant bandwidths), indeed have relatively little effect on LP-synthesised formant frequencies.

The analogous set of formant bandwidth nomograms for the 8-section LP model is shown in Figure 5.5(b), where the graphs on the left and on the right again pertain to perturbations in the parameters $a_{2n-1}$ and $b_{2n-1}$, respectively. These nomograms do support our earlier hypothesis of a distinctive relation between each odd-indexed Fourier sine parameter $b_{2n-1}$ and the corresponding bandwidth $B_n$. In particular, they show that the relation is negatively-sloped and quasi-linear. However, a comparison of Figure 5.5(b) with Figure 5.5(a) reveals that the so-called second-order influences on the bandwidths are relatively greater than those on the formant frequencies. For example, the variations in $B_1$ caused by perturbations in $a_3$ appear to be even larger in magnitude than the so-called first-order effect of $b_1$ on that bandwidth. Nevertheless, similarly to the first-order formant-frequency nomograms shown on the left in Figure
5.5(a), the first-order nomograms shown on the right in Figure 5.5(b) are each characterised by a *more strongly negative slope* than any of the other curves for the same bandwidth.

Our earlier, theoretical motivations (in Section 5.3.2.1), together with the empirical evidence just presented, strongly suggest that the odd-indexed Fourier sine components of LP-derived vocal-tract shapes are distinctively related to the formant bandwidths, whose very existence secures uniqueness in the LP-based method of inversion. Our results thereby elucidate the uniqueness property of LP-derived area-functions, for the first time in terms of their underlying shape components. Together with the earlier results presented in Section 5.3.1, we have indeed identified the intrinsic parameters of unique LP area-functions.

### 5.4 Method of Area-Function Parameterisation and Estimation

Our results presented in the previous section, were concerned with the formant dependence of LP-derived vocal-tract shapes. In particular, in Section 5.3.1 we confirmed the validity of the SM model within the LP modelling framework, and in Section 5.3.2 we then provided a theoretical motivation and empirical justifications regarding the components of LP-derived area-functions which depend on the formant bandwidths. In order to fulfill our requirements as set out in Section 5.1, in the next section we unite the newly-identified parameters of unique, LP-derived area-functions, and thereby propose a hybrid, LP-SM method of inversion. In Section 5.4.2 we then use the inversion method to evaluate our proposed parameterisation of vocal-tract shapes.

#### 5.4.1 Description

From our theoretical and empirical results presented in Section 5.3, follows our proposed extension of the SM model (cf. Equation 5.1) to include two important sets of acoustic parameters. In particular, our results concerning the formant *frequency*-dependence of LP-derived area-functions suggest that the *asymmetric* shape components $a_{2n-1}$ of the original SM model be indeed retained; and our results concerning the *bandwidth*-dependence of LP-derived area-functions suggest that the
symmetric shape components, originally represented (Schroeder, 1967; Mermelstein, 1967) by the even-indexed Fourier cosine coefficients \( a_{2n} \), be replaced with the odd-indexed Fourier sine coefficients \( b_{2n-1} \). Our proposed method of area-function parameterisation is therefore given by the following expression:

\[
\ln A(x) = \ln A_0 + \sum_{n=1}^{M/2} a_{2n-1} \cos \left( \frac{(2n-1)\pi x}{L} \right) + \sum_{n=1}^{M/2} b_{2n-1} \sin \left( \frac{(2n-1)\pi x}{L} \right),
\]

(5.18)

where \( M \) is assumed to be even, without loss of generality.

Perhaps the most distinctive feature of this new parameterisation, is that it provides a mathematically complete and minimal description of vocal-tract shapes (up to the desired degree of smoothness \( M \)). Indeed, the sine and cosine terms in Equation 5.18 form a mutually orthogonal set of basis functions which describe the vocal-tract shape in terms of the acoustically-relevant, spatial components. Furthermore, the completeness afforded by this set of basis functions does confirm both the necessity and the sufficiency of the formant frequencies and bandwidths.

As the principle of orthogonality precludes compensatory relations amongst our chosen parameters, the uniqueness property which is inherent to the LP vocal-tract model is thereby retained. However, whilst it is easily shown that all of the trigonometric terms in Equation 5.18 are mutually orthogonal, it is important to note that the symmetric components do have a non-zero mean, spatial level, and that the \( b_{2n-1} \) coefficients are therefore not strictly orthogonal to the area scaling factor \( A_0 \). This implies that in parameterising a given area-function according to Equation 5.18, the sine coefficients can only be uniquely determined if a consistent method is adopted to first determine the area scaling factor.

Towards this end, our proposed method takes advantage of the boundary conditions of each of the odd-indexed Fourier sine terms which, by definition, have zero value both at \( x = 0 \) and at \( x = L \). In particular, we first determine the antisymmetric components of the given, logarithmic area-function (by odd-indexed, discrete cosine transformation), and subtract them from the original vocal-tract shape in order to yield a primarily symmetric area function. (If the antisymmetric shape components could be determined completely by identifying an infinite number of odd-indexed Fourier cosine
coefficients, then the area function obtained after subtraction would be purely symmetric. The area scaling factor is then determined such that \(\ln A_0\) equals the mean of the logarithmic areas at the glottal and at the lip ends. Having thus defined \(A_0\), the symmetric shape parameters \(b_{2n-1}\) are found uniquely by odd-indexed, discrete sine transformation. This simple and consistent method also ensures that the first \(M/2\) of the \(a_{2n-1}\) and the \(b_{2n-1}\) parameters found for a given value of \(M\), are exactly replicated when a higher value of \(M\) is used to parameterise the same area-function.

The block diagram of Figure 5.6 shows how our proposed method of area-function parameterisation is used within the LP-based inversion method. Central to the method is the well-known set of recursive routines to convert, in turn, from formant frequencies and bandwidths (and the effective LP sampling frequency, as determined by the number of formants \(M/2\) and the assumed vocal-tract length \(L\)), first to LP autoregressive coefficients, then to LP reflection coefficients, and finally to a step-wise, LP area-function. The raw, \(M\)-section LP area-function thus obtained for a particular value of the vocal-tract length, is then parameterised according to Equation 5.18 (as described in the preceding paragraph) in order to determine the shape parameters \(\{a_{2n-1}, b_{2n-1}\} n = 1, \ldots, M/2\). This entire procedure is repeated for a range of vocal-tract lengths, and that length is finally chosen which minimises the eccentricity of the vocal-tract shape (the so-called minimal articulatory distance, or MAD) with respect to a uniform area-function. Analogously to Paige and Zue’s (1970) derivation which was based on the original SM model, it can easily be shown (e.g., using Parseval’s theorem) that the MAD criterion to be minimised when using our parameterisation method (Equation 5.18), is given by the following expression:

\[
MAD^2 = \frac{1}{L} \int_0^L [\ln A(x)]^2 dx = \frac{1}{2} \sum_{n=1}^{M/2} [a_{2n-1}^2 + b_{2n-1}^2] \tag{5.19}
\]

Consistently with the SM model, logarithmic areas are used in the articulatory distance measure in order to place greater emphasis on differences in the acoustically-salient locations of constriction along the vocal tract.

Although our method of area-function estimation illustrated in Figure 5.6 is similar to that proposed by Wakita (1977), who was indeed the first to combine the LP-based
Figure 5.6: Block-diagram of proposed area-function estimation method, valid for $L < L_{\text{max}}$.

Figure 5.7: Block-diagram of proposed area-function estimation method, to be used when an optimum vocal-tract length $L < L_{\text{max}}$ could not be found using the method of Figure 5.6.
inversion method with the vocal-tract length optimisation criterion of Paige and Zue (1970), our proposed method does extend the work of these authors. First, the restrictively coarse, step-wise representation of the vocal-tract shape yielded by the LP model, is replaced by a smooth outline from which crucial articulatory landmarks such as the place of constriction can be determined more accurately. Note that the degree of smoothness depends on the number of shape parameters used (cf. Equation 5.18), and ultimately, on the number of formants used in the inversion. We shall return to this important aspect of the proposed model in Section 5.4.2.2, where we evaluate the representational fidelity of the vocal-tract shape parameters, as a function of the degree of smoothness $M$.

As a result of the smooth representation, the computation of the MAD criterion used to optimise the vocal-tract length is no longer restricted to the entire interval $[0, L]$. Indeed, if so desired, the integral in Equation 5.19 can be numerically computed across any selected interval $[x_1, x_2]$ along the length of the vocal-tract, and different, physiologically-motivated sections of the area-function can thus be weighted differently. In that context, Yehia and Itakura (1994) determined from area-functions measured on mid-sagittal X-ray images of a single speaker of Japanese, that the two regions near the mid-length of the vocal-tract and close to the glottis, respectively, are less likely than other regions to exhibit large phonetic variations, and should therefore be given extra weight in the computation of MAD with respect to that speaker’s average area-function.

Our proposed method of inversion also overcomes the limitation of an upper bound $L_{\text{max}}$ on allowed vocal-tract lengths which, as discussed in Section 5.2.2, is inherent to all vocal-tract models which have only a limited number of sections. If the MAD criterion is not found to have a minimum for a vocal-tract length less than $L_{\text{max}}$ (which itself depends on the value of the highest formant frequency; cf. Equation 5.6), Wakita (1977) offers no alternative but to accept this upper limit as the final solution. However, it is interesting to note that Zue (1969) had already addressed this issue in his own inversion method, which was based on a completely lossless acoustic-tube model, and which therefore made use of the zeros (formants) and poles of the lip impedance function. Indeed, Zue (1969) found that the frequency of the highest singularity often
exceeded the upper spectral limit (i.e., the effective half-sampling frequency), which itself is determined by the number of sections and by the length of the vocal-tract. In order to overcome this limitation while retaining a fixed number of vocal-tract sections, he augmented the low-order poles and zeros with a number of higher-order singularities whose frequencies were chosen to be those of a uniform tube of the same length. This solution to the problem is particularly attractive, since it is founded on the theoretical result that the higher-order singularities of a non-uniform area-function converge to those of a neutral tube of the same length. It can also be shown (simply by substituting the well-known quarter-wavelength formula for the highest given formant frequency, into Equation 5.6) that if the highest singularity is equal to its neutral value for a given vocal-tract length, then that length is always less than the upper limit $L_{\text{max}}$, and the problem is thus resolved.

Our own solution to this problem in the context of the LP-based inversion method of Figure 5.6, is indeed founded on the principle of adding a higher formant to the given list of target formants. However, we transcend the simple assumption of a neutral higher formant, by invoking the very property of our resonance-based parameterisation which predicts that the additional, higher formant will primarily influence the vocal-tract shape components of correspondingly higher resolution. Whilst a neutral value for the higher formant only *aspires* towards affecting the resulting vocal-tract shape as little as possible, our method of area-function parameterisation can be used to explicitly *enforce* this outcome.

Figure 5.7 shows the block-diagram of our method of area-function estimation which is used in case the earlier method outlined in Figure 5.6 fails to minimise the MAD criterion for a vocal-tract length less than $L_{\text{max}}$. At each value of $L$ higher than that upper limit, a single, higher formant is introduced, with a frequency which is initially set to its neutral value for the given $L$, and with a bandwidth which is initially set to the mean of the target bandwidths (following our results of Section 5.3.2.1). As a result of the additional formant, the effective sampling frequency is increased by $c/L$ Hz; the number of vocal-tract sections is increased by 2; and the discrete, LP area-function is then parameterised with two more shape parameters ($a_{M+1}$ and $b_{M+1}$) than had previously been used. However, in order to maintain consistency with the vocal-
tract shapes obtained at the original (lower) shape-resolution, we insist that the higher formant must not itself contribute to additional, higher-resolution components of the resulting vocal-tract shape. In particular, as shown by the inner loop in Figure 5.7, we iteratively optimise both the frequency and bandwidth of the additional, higher formant (using steepest-descent optimisation), such that the resulting, parameterised LP area-function retains its original degree of smoothness, i.e., $F_{(M/2)+1}$ and $B_{(M/2)+1}$ are jointly optimised until $a_{M+1} = 0$ and $b_{M+1} = 0$, to within a specified tolerance — only then do we proceed to compute the MAD criterion for the area-function obtained at that particular vocal-tract length. As indicated by the outer loop in Figure 5.7, this entire procedure is repeated in search of an optimum vocal-tract length which minimises the MAD criterion, and which then yields the final, parameterised area-function.

5.4.2 Evaluation of Proposed Area-Function Parameterisation

Whilst our method of area-function parameterisation (Equation 5.18) follows naturally from our earlier derivation of the parameters of unique LP-derived area-functions, it still warrants an evaluation of its effectiveness. We shall therefore evaluate our proposed parameterisation, first (in Section 5.4.2.1) by using our hybrid method of inversion to obtain the correlations between synthetic formant and estimated shape parameters; and then (in Section 5.4.2.2) by testing the ability of our shape parameters to capture the spatial characteristics of directly-measured area-functions obtained from the literature.

5.4.2.1 Inter-parameter Correlations

We herein aim to evaluate the acoustic-phonetic relevance of our proposed set of vocal-tract shape parameters, in the context of the inversion method described in the previous section. In particular, we would like to confirm that the statistical correlation between each formant parameter and the corresponding shape parameter, is indeed as significant as suggested in our results of Section 5.3. This hypothesis is tested by first synthesising formant values from a large number of different area-functions, then applying our hybrid LP-SM method of inversion to re-estimate the shape parameters and evaluate their correlation with the synthetic formants.
First, area-functions are generated by all combinations of ternary perturbations (−0.1, 0.0, and +0.1) of the first four cosine parameters \( a_{2n-1}, \quad n = 1, \ldots, 4 \) and of the first four sine parameters \( b_{2n-1}, \quad n = 1, \ldots, 4 \), thus yielding a total of 6561 distinct shapes, with a fixed vocal-tract length \( L = 17.65 \text{ cm} \) and a normalised area scaling factor \( A_0 = 1.0 \). Assuming a nominal value for the glottal reflection coefficient \( \mu_s = 0.7 \), for which the mean of the first four bandwidths is predicted by Equation 5.17 to be \( \bar{B} = 113.5 \text{ Hz} \), the formant frequencies and bandwidths of all 6561 perturbed vocal-tract shapes are LP-synthesised. Those synthetic formants are then used to re-estimate the area-functions, using our hybrid LP-SM method of inversion.

Consistently with the SM model (and similarly to our earlier investigation in Section 5.3.1.2), correlations between vocal-tract shape and resonance parameters are computed using the relative formant frequencies \( F_n^{(\text{rel})} = (F_n - F^{(\text{neat})}) / F^{(\text{neat})} \), consistently with our results of Section 5.3.2 where we found that the shape of the LP area-function determines only the formant bandwidth pattern about the mean value, correlations are computed using the relative bandwidths \( B_n^{(\text{rel})} = (B_n - \bar{B}) / \bar{B} \). The first four, relative formant frequencies and bandwidths are then paired with each of the eight vocal-tract shape parameters in turn, and the coefficient of correlation (or Pearson’s \( r \)) is computed across all 6561 configurations.

Indeed, the high correlations along the main diagonal of Table 5.2 do provide support for a strongly linear relation between each relative acoustic resonance and the corresponding vocal-tract shape parameter. The weakest of these correlations (−0.917) summarises the strength of the linear relation between the relative second formant bandwidth and the corresponding shape parameter \( b_3 \), while the strongest correlation (−0.996) holds between the relative \( F_4 \) and the corresponding shape parameter \( a_7 \). By comparison, the most significant, off-diagonal correlation (+0.550) is that which describes the strength of the relation between the relative third formant bandwidth and the highest-indexed Fourier sine coefficient \( b_7 \). As already foreshadowed in our nomograms of Figure 5.5, the magnitude of the off-diagonal elements in Table 5.2 suggest that the influence of certain asymmetric shape parameters \( a_{2n-1} \) on the formant bandwidths, is not entirely insignificant. However, as we have already discussed, the first-order effect retains its distinctiveness owing to the negative sign of induced
perturbations in the acoustic parameters, which is here reflected in the strongly negative correlations along the main diagonal of Table 5.2.

These correlations do lend strong support to our proposed method of vocal-tract shape parameterisation. Indeed, our proposed modification and extension of the SM model is here justified quantitatively, by the strength of the one-to-one mapping between the odd-indexed Fourier sine parameters and the LP formant bandwidths. It is also interesting to note that the strong correlations obtained between the first-order pairs of formant and shape parameters, do indirectly underscore the appropriateness of the MAD criterion which was used in the inversion.

### 5.4.2.2 Representation of Directly Measured Area Functions

Our evaluation thus far has established the quasi-linear, one-to-one mapping between vocal-tract shape parameters and the LP-based formant frequencies and bandwidths. Equally relevant to this, acoustically-motivated evaluation of the parameterisation, is to consider its accuracy in merely representing area-functions obtained from direct measurements of the human vocal-tract. In that context, it is important to first evaluate the inherent smoothing of directly-measured area-functions, and then to assess the relative contribution of each parameter in representing those vocal-tract shapes.

<table>
<thead>
<tr>
<th></th>
<th>$a_1$ (rel)</th>
<th>$a_3$</th>
<th>$a_5$</th>
<th>$a_7$</th>
<th>$b_1$</th>
<th>$b_3$</th>
<th>$b_5$</th>
<th>$b_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1^{(rel)}$</td>
<td>-0.995</td>
<td>0.143</td>
<td>0.170</td>
<td>0.213</td>
<td>0.245</td>
<td>0.092</td>
<td>-0.005</td>
<td>-0.175</td>
</tr>
<tr>
<td>$F_2^{(rel)}$</td>
<td>0.150</td>
<td>-0.990</td>
<td>0.173</td>
<td>0.225</td>
<td>0.249</td>
<td>0.119</td>
<td>-0.012</td>
<td>-0.189</td>
</tr>
<tr>
<td>$F_3^{(rel)}$</td>
<td>0.174</td>
<td>0.178</td>
<td>-0.983</td>
<td>0.241</td>
<td>0.274</td>
<td>0.129</td>
<td>0.004</td>
<td>-0.224</td>
</tr>
<tr>
<td>$F_4^{(rel)}$</td>
<td>0.216</td>
<td>0.229</td>
<td>0.249</td>
<td>-0.996</td>
<td>0.334</td>
<td>0.154</td>
<td>0.008</td>
<td>-0.274</td>
</tr>
<tr>
<td>$B_1^{(rel)}$</td>
<td>0.124</td>
<td>0.528</td>
<td>0.333</td>
<td>0.113</td>
<td>-0.922</td>
<td>-0.165</td>
<td>0.174</td>
<td>0.377</td>
</tr>
<tr>
<td>$B_2^{(rel)}$</td>
<td>-0.066</td>
<td>0.041</td>
<td>0.461</td>
<td>0.135</td>
<td>-0.199</td>
<td>-0.917</td>
<td>0.299</td>
<td>0.464</td>
</tr>
<tr>
<td>$B_3^{(rel)}$</td>
<td>-0.020</td>
<td>-0.243</td>
<td>-0.062</td>
<td>0.263</td>
<td>0.236</td>
<td>0.327</td>
<td>-0.925</td>
<td>0.550</td>
</tr>
<tr>
<td>$B_4^{(rel)}$</td>
<td>-0.007</td>
<td>-0.114</td>
<td>-0.426</td>
<td>-0.357</td>
<td>0.432</td>
<td>0.421</td>
<td>0.449</td>
<td>-0.938</td>
</tr>
</tbody>
</table>

Table 5.2: Linear correlation coefficients (Pearson’s $r$) between each of the first four, relative formant frequencies and bandwidths LP-synthesised from 6561 perturbations of an 8-section, uniform area-function (with the glottal reflection coefficient fixed at a nominal value $\mu_8 = 0.7$), and each of the first eight VT-shape parameters re-estimated from those formant data using the hybrid LP-SM method of inversion described in Section 5.4.1.
The vocal-tract shapes which we shall use for this purpose, are taken from published works in which the authors have tabulated the area-functions measured using either X-ray imaging or magnetic resonance imaging (MRI) techniques. Whilst such directly-measured data are themselves prone to various, non-negligible sources of error (e.g., the necessary assumptions on the lateral width of vocal-tract sections when mapping mid-sagittal cross-dimensions to cross-sectional areas), it is generally accepted that they do provide, simply by virtue of their having been acquired by direct observation, the most accurate information that one might currently obtain on the entire shape of the human vocal-tract during vowel production.

The six, X-ray measured area-functions of Russian vowels given by Fant (1960), are perhaps the most celebrated and widely used data on vocalic area-functions, and for many years remained the only, relatively complete and easily-accessible of such data. More recent, MRI methods apparently have none of the potential hazards to which subjects are exposed during X-radiography, and in addition they can be used to reconstruct a volumetric profile of the vocal-tract airway. Nevertheless, they still suffer from a number of disadvantages, which include the requirement of repeatedly sustained articulations due to the long image-acquisition time, and the poor detection of bone structures such as the teeth, due to their low concentrations of hydrogen. Notwithstanding these limitations, a number of recent studies have used MRI to obtain invaluable area-function measurements which they have also generously published. Baer et al. (1991) provide area-functions of the four point vowels for one adult, male speaker of American English and one adult, male speaker of British English; Yang and Kasuya (1994) list area-functions of five vowels produced by two adult speakers (one male and one female) and one child, male speaker of Japanese; and Story et al. (1996) tabulate the area functions of 11 sustained vowels of a single, male speaker of American English. In addition, Beaumtemps et al. (1995) list the area-functions of the three cardinal vowels which were initially obtained from mid-sagittal X-ray images of a single, male speaker of French, and which were then refined by an optimisation procedure (applied to the mapping from cross-dimensions to areas) in order to minimise the error between the measured and resynthesised formants. In total, these five studies offer 33 directly-measured, vocalic area-functions produced by six adult, male speakers of four different
languages, including two dialects of English.

As the three area-functions of French vowels given by Beautemps et al. (1995) are originally listed with unequal section lengths, we first re-sample them by linear interpolation of the logarithmic areas in order to obtain equal-length sections not longer than the smallest original section-length (approximately 0.3 cm). Similarly, Yang and Kasuya’s (1994) five area-functions of Japanese vowels are re-sampled at equal-length intervals no greater than 0.5 cm, owing to their treatment of the area at the lip opening as a separate section of unspecified length, which we here assume to have a nominal value of 0.01 cm. Each of the 33 directly-measured (and possibly re-sampled) area-functions is then re-scaled (or normalised) to a mean logarithmic value of zero, consistently with the acoustical insignificance of the area-scaling factor in the LP vocal-tract model.

According to the well-known Nyquist sampling theorem, the number of independent parameters used to model a given area-function must not exceed the total number of vocal-tract sections. Each of the 33, re-scaled area-functions is therefore parameterised (as described in Section 5.4.1) with \( M \) equal to the highest, even integer no greater than the number of vocal-tract sections. The area-functions with the fewest number of sections (18) are those given by Baer et al. (1991); consequently, each of the 33, re-scaled and parameterised area-functions is re-expanded according to Equation 5.18, with \( M/2 \) ranging from 0 to 9, the former yielding a uniform area-function with only an area-scaling factor, and the latter yielding a smoothed vocal-tract shape with the first 18 Fourier-series terms. A quantitative measure of the goodness of fit at each degree of smoothness (integer-value of \( M/2 \)), is obtained by computing the root-mean-square (rms) error between each logarithmic area-function and its smoothed version, computed at the centre of each original, equal-length section.

The diamond symbols (joined by solid lines) in Figure 5.8 show the average rms error in smoothing the 33 directly-measured area functions, at each value of \( M/2 \). As expected, the rms error is a monotonically decreasing function of the number \( M \) of shape parameters used to represent each area-function. Our results clearly show that the first two, sine and cosine terms are the most important spatial components of directly-measured, vocalic area-functions, and that only small, incremental improvements are
gained by including terms at \( (M/2) = 3 \) and higher. This observation agrees with the conclusions drawn by Liljencrants (1971), who found it necessary to use only the first two (even-indexed) Fourier sine and cosine terms to represent X-ray measured, mid-sagittal tongue-contours mapped onto a Cartesian coordinate system.

As the degree of spatial resolution \( M/2 \) is identical to the number of formants used in the inversion, the diamond symbols in Figure 5.8 provide an optimistic benchmark for the rms error to be expected when using the inversion method. For example, if we assume that only the first four formants are available in general, the solid curve in Figure 5.8 suggests an average, lower-bound rms error of about 0.22 when comparing estimated vocal-tract shapes with so-called exact, or directly-measured area-functions. Errors in excess of this benchmark could then be ascribed to the estimation method itself, rather than the parameterisation.

In order to assess the relative contribution of antisymmetric and symmetric components of directly-measured area functions, the rms errors were also computed with respect to the smoothed vocal-tract shapes obtained by re-expanding only the Fourier cosine terms, and only the Fourier sine terms, respectively, at each value of \( M/2 \). The averaged results shown in Figure 5.8, clearly illustrate the nearly equal contribution of antisymmetric components (plus symbols, joined by dashed lines) and symmetric components (square symbols, joined by dotted lines). These results also suggest that on average, the first two cosine components \( (a_1 \text{ and } a_3) \), whose contributions are shown along the dashed lines as \( M \) is increased from 0 to 1, and from 1 to 2, respectively) and the first two sine components \( (b_1 \text{ and } b_3) \), whose contributions are shown along the dotted lines as \( M \) is increased from 0 to 1, and from 1 to 2, respectively) are the most important spatial components of the 33 area-functions.

The results shown in Figure 5.8 and discussed above, were concerned with the rms errors averaged over all of the 33 area-functions. However, as foreshadowed by Schroeder (1967) and Mermelstein (1967), individual vocal-tract shapes of certain vowels can be strongly associated with either symmetric or antisymmetric components. For example, Schroeder and Mermelstein found that the area-function of the Russian vowel \( /u/ \) given by Fant (1960), is very poorly represented by antisymmetric shape components alone, and can not therefore be adequately estimated by the SM model,
using only the formant frequencies. Their findings are confirmed by our own results (shown in Appendix E) of the individual rms profiles for each area-function, which indicate that mid- to high-, back vowels (e.g., Fant’s (1960) Russian /u/; Baer et al.’s (1991) British English /u/ and American English /u/; Yang and Kasuya’s (1994) Japanese /u/ and /o/; Beautemps et al.’s (1995) French /u/; and Story et al.’s (1996) American English /o/, /a/, and /u/) most consistently exhibit vocal-tract shape symmetry, owing presumably to the location of the linguo-velar constriction at approximately midway along the length of the vocal-tract.

By contrast, the two point vowels /i/ and /a/, which are usually associated with extremes of lingual articulation in the high-front and in the low-back parts of the vocal tract, are the most consistently antisymmetric of the directly-measured area-functions.
(except for Beautemps et al.’s (1995) French /a/, which admittedly has a more symmetric shape, owing to the much smaller and decreasing areas towards the lip end).

From our quantitative evidence may be inferred the physiological point of view that asymmetry in vocal-tract shapes implies conservation of the mass of the tongue body — for example, when the tongue moves forwards and upwards to form a constriction for a high-front vowel, the pharyngeal cavity is endowed with a proportionately larger volume; and when it moves backwards and downwards to form a constriction for a low-back vowel, larger areas are naturally attained in the front, oral cavity.

Our results can also be interpreted from an acoustical point of view, as a consequence of the close relationships which we have shown to exist between antisymmetric shape components and the formant frequencies, and between symmetric shape components and the formant bandwidths. Indeed, the relative dominance of the first two Fourier-series shape components, does confirm the relative importance of the first two formants in vowel production. The significant contribution of the two antisymmetric shape parameters $a_1$ and $a_3$, in particular, provides indirect proof (gleaned here only from results of shape parameterisation) that $F_1$ and $F_2$ are indeed the most important acoustic determinants of the phonetic identity of vowels. At the same time, however, the significance of the symmetric shape parameters in representing directly-measured area-functions, is consistent with their importance in securing uniqueness in LP-derived vocal-tract shapes.

### 5.5 Evaluation of Estimation Method

Evaluation of any proposed method of inversion has always been a major weakness in studies concerned with acoustic-to-articulatory mapping. Whilst confidence is usually raised if acoustic and articulatory data are simultaneously measured, existing sources of error in directly measured articulatory data (some of which were discussed earlier, in Section 5.4.2.2) do suggest that they should not be given carte blanche as the ultimate reference. On the other hand, there seems to have been some over-reliance on indirect evaluations, such as perceptual or spectrographic fidelity of resynthesised speech — it has rightly been asserted (e.g., Sondhi, 1979) that, owing mainly to the problem of non-uniqueness, resynthesis alone is not sufficient for evaluating estimated articulatory data.
In this vein, perceptual judgements are probably even more inconclusive, owing to the suggested “many-to-one relationship between articulatory/acoustic and perceptual targets” (Gay et al., 1991, p.445) for vowel sounds. These limitations notwithstanding, directly-measured area functions are admittedly the best reference that we have regarding data on real vocal-tract shapes. However, owing to technical obstacles, simultaneously-measured acoustic data are rare, and even when they are documented, the formant bandwidths are invariably missing. It would therefore seem expedient to evaluate our method of inversion first using acoustic data obtained by synthesis which, although admittedly biased towards a more favourable assessment, does afford a systematic definition of the various sources of error. To this end, in Section 5.5.2 we will take as a point of departure the 33 published, directly-measured vocal-tract area-functions which we used earlier (in Section 5.4.2.2), and impress certain assumptions which enable us to evaluate our hybrid LP-SM method of inversion first under model-matched (i.e., LP-matched), then under presumably more realistic conditions.

In Section 5.5.3 we then proceed to estimate vocal-tract shapes from the measured formant data of our four adult, male speakers of Australian English. In particular, we will first use our measurements of the first four formant frequencies and bandwidths, to quantify the formant-induced variability of LP-estimated area-functions, and thus arrive at a prescription for dealing with the large amounts of variability normally encountered in measured bandwidths. We will then consider the well-known, but ill-treated problem of the incompatibility of measured formant bandwidths with those ideally required by the LP vocal-tract model.

However, a necessary requirement in the forthcoming evaluation is to be able to quantify the variability in estimated vocal-tract shapes. As area-functions are generally of different lengths, and as variations in the horizontal position of the lips and in the vertical position of the larynx preclude them from being considered as fixed articulatory landmarks, the problem of area-function alignment is highly non-trivial; nor has it been addressed to a sufficient degree in the literature. In Section 5.5.1 we therefore propose a method of aligning parameterised area-functions, which we shall then use both in our subsequent evaluation of our estimation method, and also later in our articulatory
investigation of the speech-speaker dichotomy in Chapter 6.

5.5.1 Inter-repetition Alignment of Area-Functions

As already discussed in our review of the literature (in Chapter 2), the problem of directly comparing two or more vocal-tract shapes and quantifying their differences, has rarely been addressed in previous studies. A major obstacle which has impeded progress in that regard, is presumably the absence of an accepted method of aligning vocal-tract area-functions (whether directly measured or estimated from acoustics) such as to render physically meaningful comparisons. The problem is exacerbated by the well-known tendency for the position of the lips and of the larynx to vary from one articulatory configuration to another, thus prohibiting the use of either end of area-functions as an invariant positional reference. As the overall lengths of different area-functions are also different in general, one is faced with the non-trivial problem of how best to align area-functions prior to quantifying their shape-related differences.

One might suppose that if certain, fixed articulatory landmarks could be located along each area function, then they might be aligned by piecewise-linear expansion or contraction along the length dimension, anchored at those landmarks. Unfortunately, this attractive solution is not only unrealistic in the context of area-function estimation, but would also be quite difficult and laborious to achieve with a high degree of precision even with direct articulatory measurements. Previous approaches to this problem include Harshman et al.’s (1977) elaborate definition of a speaker-specific, mid-sagittal reference grid; Wood’s (1979) alignment of area functions (measured from X-ray films) by anchoring them at the single point of reference defined by the position of the central incisors; and Högberg’s (1995) inter-gender normalisation of the total lengths of X-ray measured area functions, by a piecewise-linear rescaling which relies on the definition of the boundary between the oral and the pharyngeal parts of the vocal tract. Clearly, such methods cannot be applied to estimated vocal-tract shapes, which lack the convenience of physiological landmarks.

However, if one requires to quantify the variability, or the dispersion in a group of area-functions which are variants of the same, basic vocal-tract shape, then one might
take advantage of the fact that those shapes are assumed to share a common set of articulatory characteristics. This condition may arise, for example, when area-functions are to be compared across a number of analysis frames and repetitions of a given vowel of a single speaker. One way of overcoming the lack of a fixed frame of reference, is then to allow each area function a positional degree of freedom along the length axis, subject to the constraint of a minimal amount of variation about the typical, or prototype vocal-tract shape which they all are assumed to embody. The problem of aligning the given set of (presumed) phonetically equivalent area functions, thus reduces to a minimisation of the root-mean-square (rms) dispersion of the vocal-tract shapes relative to each other.

As initially proposed and demonstrated by Clermont (1991; 1993) who successfully time-aligned repetitions of formant-contours of diphthongs, the easiest approach (shown schematically in Figure 5.9) is to slide each shape (formant-contour or, in the present study, vocal-tract area-function) along the length axis, until the rms difference is minimised with respect to a so-called reference shape, which itself is
chosen to be the longest one of the set. As the shapes are, by definition, not too
dissimilar, a global minimum in the rms error can be found by sliding each, so-called test
area-function, only within certain limits which are defined by so-called relaxation
intervals placed at both ends of the reference area-function.

In practice, this is achieved by first expanding the Fourier series terms in Equation
5.18 in order to obtain the smoothed versions of the parameterised area-functions
themselves, sampled at small, equal-length intervals $\Delta x$. The area-function with the
longest length is then selected as a reference, and a sufficient number of sections is
allowed at either end to accommodate the relaxation intervals, each of length $l_{rel}$. Each
of the test area-functions is then shifted within these outer limits, and at each step the
rms distance is computed between the test and reference, logarithmic area-functions
(consistently with the SM model), over that part of the overall length which is common
to both (the mutually-overlapping region, as hereafter referred to). The global
minimum in the rms error is thus found to within an accuracy of $\pm \frac{1}{2} \Delta x$ for each area-
function.

As a result of the procedure outlined above, each of the parameterised vocal-tract
shapes is endowed with an extra parameter which identifies its position along the $x$-axis,
relative to a fixed frame of reference defined by the longest area-function in that set. A
so-called prototype area-function can then be found simply by computing the average of
the (logarithmic) areas at every interval $\Delta x$. The total length of the prototype is
determined by the mean length of the area-functions in the set; and its end-points are
located equi-distantly about the mid-point of the overall mutually-overlapping (MOL)
region. Finally, the prototype area-function thus found, can be parameterised according
to Equation 5.18, and regarded as the basic vocal-tract shape of which the given set of
area-functions were assumed to be variants.

As we shall see in Chapter 6, the procedure just outlined will prove to be very
useful in reducing the multi-frame and multi-repetition area-functions, to yield a single,
prototype vocal-tract shape for each vowel of each speaker. In the following sections,
we shall use it primarily to quantify the variability in vocal-tract shapes estimated under
various conditions, by retaining the minimum rms error yielded after each alignment.
5.5.2 Re-estimation of Directly Measured Area-Functions

The first part of our evaluation concerns re-estimation of vocal-tract area-functions which have already been directly measured using X-ray imaging or MRI techniques — in particular, the 33 area-functions obtained from the literature and used earlier in Section 5.4.2.2. The formant data required as input to our hybrid LP-SM method of inversion, will be synthesised first under model-matched conditions (in Section 5.5.2.1), then under more realistic conditions (in Section 5.5.2.2). Our method of inter-repetition alignment will be used to quantify the differences between the original and re-estimated area-functions, thereby assessing the detrimental influences of using model-mismatched acoustic data. We then appraise (in Section 5.5.2.3) the effectiveness of using such model-based results in correcting the (presumed) “more realistic” formant data prior to inversion.

5.5.2.1 Model-matched Conditions

The LP model which lies at the heart of our inversion method, makes certain unrealistic assumptions (as discussed earlier in Section 5.2.2), the two most drastic of which are the following: (i) the number of equal-length vocal-tract sections is equal to twice the number of formants considered; and (ii) the only source of losses in the vocal-tract is a frequency-independent, resistive element at the glottal end. We herein conform to the first of these assumptions, by representing each of the 33 directly-measured area-functions in terms of 8 equal-length sections, obtained by parameterisation (with \( M = 8 \) in Equation 5.18) and re-expansion of the areas at the centre of each section (i.e., effectively by interpolating the smoothed area-function at equal-length intervals). Assuming a nominal value \( \mu_g = 0.8 \) for the glottal reflection coefficient, the first four formant frequencies and bandwidths of each of the smoothed, 8-section area-functions are then synthesised using a vocal-tract acoustic-simulation model (see Appendix C for details) with all losses removed except for the glottal resistance. The synthetic formants thus obtained under model-matched conditions (and listed in Table 5.3), are then used to re-estimate each of the area-functions, using our hybrid LP-SM method of inversion described earlier in Section 5.4.1.
<table>
<thead>
<tr>
<th>Study</th>
<th>Vowel</th>
<th>Formant Frequency (Hz)</th>
<th>Formant Bandwidth (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$F_1$</td>
<td>$F_2$</td>
</tr>
<tr>
<td>Fant (1960)</td>
<td>/i/</td>
<td>232.4</td>
<td>2271.5</td>
</tr>
<tr>
<td></td>
<td>/e/</td>
<td>440.5</td>
<td>2025.2</td>
</tr>
<tr>
<td></td>
<td>/æ/</td>
<td>708.3</td>
<td>1232.3</td>
</tr>
<tr>
<td></td>
<td>/ɔ/</td>
<td>564.7</td>
<td>1028.8</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>298.2</td>
<td>665.6</td>
</tr>
<tr>
<td></td>
<td>/i/</td>
<td>294.6</td>
<td>1897.3</td>
</tr>
<tr>
<td>Baer et al. (1991)</td>
<td>/i/</td>
<td>271.6</td>
<td>2144.9</td>
</tr>
<tr>
<td>(Speaker TB)</td>
<td>/æ/</td>
<td>641.7</td>
<td>1529.3</td>
</tr>
<tr>
<td></td>
<td>/ɑ/</td>
<td>524.3</td>
<td>1165.8</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>302.4</td>
<td>1335.7</td>
</tr>
<tr>
<td>Baer et al. (1991)</td>
<td>/i/</td>
<td>282.9</td>
<td>2230.1</td>
</tr>
<tr>
<td>(Speaker PN)</td>
<td>/æ/</td>
<td>694.0</td>
<td>1656.9</td>
</tr>
<tr>
<td></td>
<td>/ɑ/</td>
<td>600.5</td>
<td>1337.6</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>318.3</td>
<td>1806.5</td>
</tr>
<tr>
<td>Yang &amp; Kasuya</td>
<td>/i/</td>
<td>284.2</td>
<td>2103.7</td>
</tr>
<tr>
<td>(1994)</td>
<td>/ɛ/</td>
<td>541.7</td>
<td>1844.9</td>
</tr>
<tr>
<td></td>
<td>/ʌ/</td>
<td>710.9</td>
<td>1211.5</td>
</tr>
<tr>
<td></td>
<td>/ɑ/</td>
<td>578.9</td>
<td>1067.2</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>415.0</td>
<td>1581.2</td>
</tr>
<tr>
<td>Beautemps et al.</td>
<td>/i/</td>
<td>232.9</td>
<td>2013.0</td>
</tr>
<tr>
<td>(1995)</td>
<td>/æ/</td>
<td>522.4</td>
<td>1338.7</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>362.5</td>
<td>1030.4</td>
</tr>
<tr>
<td>Story et al.</td>
<td>/i/</td>
<td>236.4</td>
<td>2441.2</td>
</tr>
<tr>
<td>(1996)</td>
<td>/ɪ/</td>
<td>501.3</td>
<td>2189.5</td>
</tr>
<tr>
<td></td>
<td>/ɛ/</td>
<td>670.4</td>
<td>2196.3</td>
</tr>
<tr>
<td></td>
<td>/æ/</td>
<td>761.7</td>
<td>1971.9</td>
</tr>
<tr>
<td></td>
<td>/ʌ/</td>
<td>754.3</td>
<td>1379.8</td>
</tr>
<tr>
<td></td>
<td>/ɑ/</td>
<td>835.7</td>
<td>1232.0</td>
</tr>
<tr>
<td></td>
<td>/ɔ/</td>
<td>673.7</td>
<td>1156.1</td>
</tr>
<tr>
<td></td>
<td>/ʊ/</td>
<td>477.3</td>
<td>1019.7</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>487.2</td>
<td>1003.8</td>
</tr>
<tr>
<td></td>
<td>/ʊ/</td>
<td>281.0</td>
<td>1395.8</td>
</tr>
<tr>
<td></td>
<td>/ɔ/</td>
<td>570.3</td>
<td>1662.8</td>
</tr>
</tbody>
</table>

Table 5.3: Formant frequencies and bandwidths of 33 directly-measured area-functions, synthesised using a transmission-line analog of the vocal-tract (see Appendix B) with a glottal resistance as the only source of loss. Each area-function is first parameterised with $M = 8$ shape parameters, which are then used in Equation 5.18 to obtain 8 equal-length sections. The glottal reflection coefficient is fixed at a nominal value $\mu_8 = 0.8$. 240
As the LP-based inversion is theoretically reversible, the 8-section area-functions yielded by constraining the inversion to use the original vocal-tract lengths are almost identical to the original, 8-section area-functions. Indeed, the average rms error obtained after alignment (with $\Delta x = 0.05 \text{ cm}$ and $l_{dz} = 2.0 \text{ cm}$), was found to be only 0.04. However, the MAD criterion can only be tested by allowing it to determine the optimum length of each area-function under these, so-called model-matched conditions.

In particular, the method depicted in Figure 5.6 is first used in search of a minimum in the MAD criterion, for a range of vocal-tract lengths starting from 12cm and incrementing in steps of 0.4 cm up to the limit $L_{\text{max}}$, which is determined by the highest (the fourth) formant frequency (cf. Equation 5.6 in Section 5.2.2). A second search is then carried out over a much smaller range of vocal-tract lengths which brackets the identified minimum in the MAD criterion, to determine the optimum length at a finer resolution of 0.05 cm. Of the 33 area-functions considered, only two (Fant’s (1960) Russian /u/ and Story et al.’s (1996) American English /u/) were found not to have a minimum in the MAD criterion for a length less than their respective $L_{\text{max}}$. The optimum lengths of these area-functions were therefore obtained using the method depicted in Figure 5.7, with an additional, fifth formant used to overcome the $L_{\text{max}}$ limitation (as described in Section 5.4.1).

In Figure 5.10 we illustrate typical curves of MAD versus $L$ obtained in finding an optimum vocal-tract length, shown here only for the six area-functions of Fant (1960). Our results are indeed comparable to the curves obtained by Paige and Zue (1970) for the same area-functions, even though their method of inversion was somewhat different from ours. Similarly to their results, Figure 5.10 indicates that the high-front vowel /i/ is the most eccentric (relative to a uniform tube) of the six vocal-tract shapes, even after optimisation of the vocal-tract length to minimise eccentricity; the optimised area-function of the mid-front vowel /e/ is not only the shortest, but also the least eccentric of the six vocal-tract shapes; the re-estimated area-function of the lip-rounded vowel /u/ is found to be the longest. However, contrary to the results of Paige and Zue (1970, Table III) and of Wakita (1977, Figure 6), who both slightly underestimated the length of this high-back vowel, our LP-based method clearly overestimates its length, yielding the largest, positive error (13.6 %) amongst the re-estimated lengths of the 33 area-
Figure 5.10: Curves of MAD versus $L$, obtained in determining the optimum vocal-tract length of each of Fant’s (1960) six Russian vowels. The vocal-tract shapes are re-estimated using the method given in Section 5.4.1 (cf. Figures 5.6 and 5.7), with the MAD criterion defined in Equation 5.19. The first four, synthetic formant frequencies and bandwidths (cf. Table 5.3) are used as input to the inversion method.

Figures 5.11: Scatter-plot of re-estimated versus original vocal-tract lengths, for the 33 directly-measured area-functions obtained from the literature. The vocal-tract shapes are re-estimated from the first four formant frequencies and bandwidths (cf. Table 5.3) synthesised under model-matched conditions. Solid line: 45-degree line drawn for reference only. Dashed line: result of a linear regression analysis (correlation coefficient 0.82; cf. Equation 5.20).
functions.

The scatter-plot in Figure 5.11 shows the relation between the re-estimated and original vocal-tract lengths for all 33 area-functions considered. The distribution of the scatter suggests that our method of area-function estimation tends to underestimate the lengths of originally shorter area-functions, and to overestimate the lengths of originally longer area-functions, thus exaggerating the normal variation in vocal-tract length from vowel to vowel. Indeed, a linear regression analysis of the 33 pairs of vocal-tract lengths, yields a correlation coefficient of 0.82, and the following relation:

$$L_{\text{re-estimated}} = 1.46L_{\text{original}} - 8.06,$$

(5.20)

which is shown by the dashed line in Figure 5.11. It is interesting to note that the line of best fit intersects the solid line (which represents an ideal condition where the re-estimated and the original lengths are identical) at about $L = 17.5$ cm, which is known to be approximately the average vocal-tract length of an adult male.

The obvious exaggeration of inter-vowel variations in vocal-tract length notwithstanding, it is important to note that the vocal-tract shapes themselves are re-estimated remarkably well. Indeed, as illustrated by the dashed curves superimposed on the graphs in Appendix F, articulatory features such as the place of lingual constriction, and the place of labial constriction for the lip-rounded vowels, are preserved fairly accurately. For example, whilst the total length of the re-estimated area-function of Fant’s (1960) /u/ is clearly overestimated by the largest amount (as discussed earlier), the re-estimated vocal-tract shape itself preserves (after alignment) not only the places of linguo-velar and labial constriction at approximately 7-8 cm and 19 cm from the original position of the glottis, respectively, but also retains the shape and volume of the back and the front cavities. The overestimated length appears to be therein manifested mainly by an exaggeration in the amount of lip-protrusion, and also partly by the position of the larynx, which is effectively lowered by 0.55 cm compared with original. Similarly, although the re-estimated length of Story et al.’s (1996) area-function for /a/ is underestimated by the largest amount (~15.2 %), its shape retains the main place of constriction at approximately 6 - 6.5 cm from the glottis, and appears to be adversely affected mainly towards the front part of the oral cavity.
Figure 5.12: Root-mean-square (rms) errors computed using the alignment procedure (cf. Section 5.5.1), between vocal-tract shapes re-estimated from the first four formants synthesised under various conditions, and the shapes which represent each original area-function with the same degree of smoothness ($M/2=4$), for each of 33 directly-measured area-functions obtained from the literature. 

Diamond symbols, joined by solid lines: 8-section area-functions; LP synthesis (cf. Table 5.3). 

Plus symbols, joined by dashed lines: original section-lengths and areas; a lossy vocal-tract model (cf. Table 5.4).
In drawing conclusions regarding the appropriateness of the MAD criterion for estimating the vocal-tract length, we are compelled, in light of the evidence discussed above, not only to compare the re-estimated and the original lengths, as earlier studies have done, but also to quantify the dissimilarity between the original and the re-estimated vocal-tract shapes. As discussed earlier in Section 5.2.2, quantification of differences between raw, step-wise LP area-functions of different lengths is non-trivial (and indeed has apparently never been attempted!), owing partly to the fact that the original and the re-estimated section boundaries would no longer coincide. By contrast, our method of area-function parameterisation yields a smoothed representation of each area-function, and the alignment method described in Section 5.5.1 can then be used to quantify the differences between original and re-estimated vocal-tract shapes.

The rms distances thus computed after alignment of each re-estimated vocal-tract shape and the smoothed, parameterised version of the corresponding, original area-function, are shown by the diamond symbols (joined by solid lines) in Figure 5.12. Whilst they all are, as expected, numerically larger than the mean rms distance (0.04) computed earlier for the vocal-tract shapes re-estimated using their original lengths, the mean of the 33 distances (0.12) is still lower than the mean rms error (0.22) incurred in representing those area-functions using the same number of shape parameters (as obtained in Section 5.4.2.2). The largest, anterior shift required to align the pairs of vocal-tract shapes was found to be 0.65 cm, for Story et al.’s (1996) American English /a/, the length of which was the most underestimated; the rms distance computed for that area-function (0.45) is also the largest, and is mainly attributable to the inaccurate re-estimation of the size of the oral cavity and of the lip opening area. The largest alignment-shift towards the glottal end was found to be –0.55 cm, for Fant’s (1960) Russian /u/, the length of which was the most overestimated. The remaining vocal-tract shapes were found to require shifts along the length-axis by amounts intermediate between those two extremes, and they generally also yielded lower, rms distances. It is encouraging, therefore, to note that the close resemblance of the re-estimated shapes to the original area-functions (as shown in the graphs of Appendix F), is quantitatively confirmed by the rms errors yielded by our automatic method of alignment.

The results presented above, collectively provide quantitative support for the
inversion method used under model-matched conditions, and with only a finite number of formants. Thus, they also lend credence to the MAD criterion, by which the vocal-tract length (and effectively the vocal-tract shape) is determined. However, in reality, the formant frequencies and bandwidths measured from the acoustic speech signal are not “matched” with the LP vocal-tract model. We therefore continue our evaluation of the inversion method in the next section, using somewhat more realistic, “model-mismatched” conditions.

5.5.2.2 More Realistic Conditions

In order to introduce more realistic conditions in our evaluation of the inversion method, we herein relax the constraints arising from two critical assumptions underlying the LP vocal-tract model referred to in the previous section. The first of those assumptions concerns the rarely acknowledged fact, that the formants synthesised using any type of transmission-line analog of the vocal-tract, depend to a certain extent on the number of sections used to represent a given area-function. Indeed, it was to avoid mismatch in synthesised formants which led Wakita (1977, 1979) to first reduce Fant’s (1960) area-functions to 8 sections prior to using the LP method (with four formants) to re-estimate those vocal-tract shapes. As it is well known (e.g., Fant, 1960) that a more accurate acoustic simulation of the vocal-tract is afforded by using a greater number of sections, we introduce more realistic conditions (and thus a model-mismatch) by synthesising the formants using the original section-lengths and areas of the 33 area-functions used previously.

The second, crucial assumption concerns the distribution of acoustic energy losses in the vocal-tract. More realistic conditions are simulated by using a transmission-line analog which does include not only the glottal resistance (once again assuming a nominal value of 0.8 for the glottal reflection coefficient), but also a glottal inductance; a viscosity factor, heat-conduction and wall-vibration losses at each section; and a radiation impedance at the lips (see Appendix C for a more detailed description). The first four formant frequencies and bandwidths thus synthesised (and listed in Table 5.4) are then used to re-estimate each of the 33 area-functions using our hybrid LP-SM method of inversion.
<table>
<thead>
<tr>
<th>Study</th>
<th>Vowel</th>
<th>Formant Frequency (Hz)</th>
<th>Formant Bandwidth (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$F_1$</td>
<td>$F_2$</td>
</tr>
<tr>
<td>Fant (1960)</td>
<td>/i/</td>
<td>293.1</td>
<td>2311.5</td>
</tr>
<tr>
<td></td>
<td>/e/</td>
<td>461.1</td>
<td>1997.6</td>
</tr>
<tr>
<td></td>
<td>/a/</td>
<td>686.6</td>
<td>1108.4</td>
</tr>
<tr>
<td></td>
<td>/o/</td>
<td>546.4</td>
<td>888.7</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>301.3</td>
<td>631.3</td>
</tr>
<tr>
<td></td>
<td>/i/</td>
<td>339.8</td>
<td>1524.3</td>
</tr>
<tr>
<td>Baer et al. (1991)</td>
<td>/i/</td>
<td>322.3</td>
<td>2185.3</td>
</tr>
<tr>
<td>(Speaker TB)</td>
<td>/e/</td>
<td>643.8</td>
<td>1442.5</td>
</tr>
<tr>
<td></td>
<td>/a/</td>
<td>540.2</td>
<td>1089.3</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>341.4</td>
<td>1097.0</td>
</tr>
<tr>
<td>Baer et al. (1991)</td>
<td>/i/</td>
<td>333.2</td>
<td>2293.1</td>
</tr>
<tr>
<td>(Speaker PN)</td>
<td>/e/</td>
<td>695.5</td>
<td>1595.0</td>
</tr>
<tr>
<td></td>
<td>/a/</td>
<td>622.8</td>
<td>1269.6</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>351.5</td>
<td>1465.4</td>
</tr>
<tr>
<td>Yang &amp; Kasuya (1994)</td>
<td>/i/</td>
<td>329.3</td>
<td>2082.5</td>
</tr>
<tr>
<td></td>
<td>/e/</td>
<td>545.2</td>
<td>1598.8</td>
</tr>
<tr>
<td></td>
<td>/a/</td>
<td>670.4</td>
<td>1080.9</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>508.4</td>
<td>858.7</td>
</tr>
<tr>
<td></td>
<td>/i/</td>
<td>540.2</td>
<td>1089.3</td>
</tr>
<tr>
<td>Beaupre et al. (1995)</td>
<td>/i/</td>
<td>292.9</td>
<td>2070.4</td>
</tr>
<tr>
<td></td>
<td>/e/</td>
<td>518.7</td>
<td>1304.9</td>
</tr>
<tr>
<td></td>
<td>/a/</td>
<td>358.8</td>
<td>867.4</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>508.4</td>
<td>858.7</td>
</tr>
<tr>
<td>Story et al. (1996)</td>
<td>/i/</td>
<td>290.1</td>
<td>2521.1</td>
</tr>
<tr>
<td></td>
<td>/e/</td>
<td>648.5</td>
<td>2009.0</td>
</tr>
<tr>
<td></td>
<td>/a/</td>
<td>730.9</td>
<td>1840.2</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>692.2</td>
<td>1302.1</td>
</tr>
<tr>
<td></td>
<td>/i/</td>
<td>808.9</td>
<td>1147.9</td>
</tr>
<tr>
<td></td>
<td>/a/</td>
<td>632.6</td>
<td>1069.7</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>403.7</td>
<td>868.4</td>
</tr>
<tr>
<td></td>
<td>/e/</td>
<td>407.3</td>
<td>943.1</td>
</tr>
<tr>
<td></td>
<td>/a/</td>
<td>315.1</td>
<td>1126.0</td>
</tr>
<tr>
<td></td>
<td>/u/</td>
<td>535.2</td>
<td>1572.9</td>
</tr>
</tbody>
</table>

Table 5.4: Formant frequencies and bandwidths of 33 directly-measured area-functions, synthesised using a transmission-line analog of the vocal-tract (see Appendix C) which includes glottal impedance, viscosity, heat-conduction, wall-vibration, and lip-radiation losses. The original, published section-lengths and areas are used, and the glottal reflection coefficient is fixed at a nominal value $\mu_{glott} = 0.8$. 

247
The re-estimated area-functions are shown in Appendix F (dotted curves), superimposed on each of the original area-functions (step-wise, solid lines) from which the formants were synthesised, and the area-functions (dashed curves) re-estimated in the previous section under the model-matched conditions. It is immediately apparent from these graphs, that despite the deliberate mismatch in the number of sections and in the vocal-tract acoustic model used to synthesise the formants, the re-estimated shapes still capture most of the essential, articulatory features such as constriction locations and cavity sizes, often with remarkable fidelity. Indeed, there do not appear to be any catastrophic errors, as for example a front vowel having a back place of constriction.

On the other hand, a number of area-functions are not so well re-estimated. The worst cases are the area-functions of the vowel /u/, for which a comparison of Tables 5.3 and 5.4 reveals not only a change in the bandwidth pattern, but also a drop, often more than 200Hz, in the second formant frequency; this has been found to be caused primarily by the mismatch in the number of vocal-tract sections. However, this does not appear to affect Fant’s (1960) Russian /u/ which already has quite a backward place of lingual articulation; rather, it affects mainly the remaining, more fronted configurations of /u/, and similarly the fronted /ɨ/ of Fant (1960). It is interesting to note that some of these area-functions are amongst those found earlier (in Section 5.4.2.2) to have a distinctively symmetric shape.

Our qualitative observations of the superimposed area-functions are confirmed in Figure 5.12, where the rms errors obtained after alignment are shown by the plus symbols (joined with dashed lines), superimposed with those obtained earlier. Indeed, the largest increases in rms error compared with the previous, model-matched results, occur for /u/ of Story et al. (1996) and of Baer et al.’s (1991) speaker PN. Not unexpectedly, rms errors generally increase as a result of the model mismatch. However, in a number of cases the rms error is only slightly worse than before (Fant’s (1960) /u/, Yang and Kasuya’s (1994) /u/ and /e/, and Story et al.’s (1996) /ʌ/, /ɑ/, and /ɔ/); and most intriguingly, the rms error obtained for Yang and Kasuya’s /ɑ/ is actually lower than that obtained under model-matched conditions.

The results discussed above can be summarised in terms of the mean of the rms errors for all 33 area-functions (0.37) which, although greater than the mean rms error
(0.22) found in representing those area-functions with the same number of parameters, might still be regarded favourably, considering the large number of fairly accurately re-estimated vocal-tract shapes shown in Appendix F. Indeed, not only do those superimposed area-functions show little evidence of gross errors, they also demonstrate that the inversion method is likely to succeed in capturing the main articulatory features of the majority of the vocalic area-functions considered, to within the spatial resolution afforded by the first $M = 8$ shape parameters, and using only the first four formant frequencies and bandwidths.

5.5.2.3 Model-based Formant Correction

Clearly, re-estimation of directly-measured area-functions using formants synthesised under more realistic conditions than afforded by the LP vocal-tract model, generally yields worse results. In particular, we have found (in Section 5.5.2.2) that the mean rms error between original and re-estimated vocal-tract shapes under model-mismatched conditions, is about three times larger than that obtained (in Section 5.5.2.1) under model-matched conditions. Nearly two decades ago, Wakita (1979) proposed to combat this problem by using a so-called “formant conversion chart” to compensate for the effects on the formant frequencies, of the LP model’s incompleteness in regard to vocal-tract losses. However, his conversion chart apparently does not include the effects of using different numbers of vocal-tract sections; nor has there appeared a conversion chart for the formant bandwidths.

A more complete version of Wakita’s conversion chart is afforded by the formants which we have already synthesised from the 33 directly-measured area-functions, first under model-matched, then under more realistic conditions. In particular, the 4 formant frequencies listed earlier in Tables 5.3 and 5.4 are plotted along the ordinate and the abscissa, respectively, in Figure 5.13(a); similarly the 4 formant bandwidths, which are plotted in Figure 5.13(b). The solid curve shown superimposed in Figure 5.13(a), is a cubic-polynomial fit to the 132 pairs of data points in the logarithmic-frequency domain, and is described as follows:

$$
\ln \left( F_n^{(LP)} \right) = 0.0494 \ln \left( F_n^{(FL)} \right) - 1.14 \ln \left( F_n^{(FL)} \right)^2 + 9.711 \ln \left( F_n^{(FL)} \right) - 21.8
$$

(5.21)
Figure 5.13: Formant frequency (a) and bandwidth (b) conversion charts (more complete version of the frequency chart first proposed by Wakita, 1979). The first four formants of each of the 33 directly-measured area-functions obtained from the literature, are synthesised using first the LP vocal-tract model with the glottal reflection coefficient fixed at a nominal value $\mu_{\text{glott}} = 0.8$ (ordinate, see Table 5.3), then a lossy transmission-line analog of the vocal-tract (abscissa, see Table 5.4). The solid curve in (a) is the best cubic fit to the set of logarithmic-frequency data, and is described by Equation 5.21 in the text.
As pointed out by Wakita (1979), the wall-impedance effects in the low-frequency, \( F_1 \) range appear to cause the most substantial deviation from the 45º straight (dashed) line. In addition, our more complete, formant frequency conversion chart highlights the extent of model mismatch caused by the combination of vocal-tract losses and number of sections, which appear mainly to affect the \( F_2 \) range.

In contrast with the formant frequencies, our formant bandwidth conversion chart shown in Figure 5.13(b) does not suggest any regularity which might be exploited by fitting the data with a function. On the contrary, the data appear to be spread across regions of the chart which lie far away from the ideal, 45º line. Taken one bandwidth at a time, the \( B_1 \) and the \( B_2 \) data points (diamond and plus symbols, respectively) do appear to form an elongated cluster which may be fit with a straight line of large, positive slope; the \( B_4 \) data points (cross symbols) might be considered to form a cluster which is elongated horizontally, and which might therefore be fit with a straight line having very little variation in the LP-based values along the ordinate. On the other hand, the \( B_3 \) data points (square symbols) appear to be scattered across a large region of the chart, with no apparent regularity. Clearly, formant bandwidth correction is not as amenable a task as formant frequency correction.

Before becoming too despondent over these problematic results, it is worth noting that they are purely model-based (i.e., they are based entirely on formants synthesised using either the LP or a more lossy vocal-tract model). In this vein, we have already questioned (in Section 5.2.2) the viability of using model-based correction procedures, when it remains unknown to what degree those models can be trusted to be accurate or realistic. Indeed, the fidelity of even a fully-lossy vocal-tract model which does include many of the acoustic and aerodynamic properties missing from the LP model, has not been firmly established. Perhaps we should therefore not expect model-based formant correction procedures, even if they could be formulated, to significantly and consistently improve the plausibility of LP-derived vocal-tract shapes by rendering them somehow more “realistic” than they already are.

In that context, it is highly instructive to enquire whether those five studies from which we have taken the 33 directly-measured area-functions, have also reported results concerning the discrepancy between measured and re-synthesised formant data. Those
studies have indeed published two sets of formant frequency values based, respectively, on acoustic measurements (from recordings made either at the time of the articulatory measurements or, more commonly, on a separate occasion) and vocal-tract simulations using the measured area-functions themselves. Fant (1960, Table 2.31-1, p.109) used a digital computer (BESK) implementation of a 20-section vocal-tract model, and found that the “average deviation of calculated data from spoken data is of the order of 5 per cent in $F_2$ and $F_3$ and 10 per cent in $F_1$”; Baer et al. (1991, Table III, p.813) synthesised waveforms from their measured area-functions using a variant of Kelly and Lochbaum’s (1963) model which includes non-ideal terminations at the lips and at the glottis, and found that the “formant frequencies of the subjects’ utterances differ significantly from the computed resonances of the area functions”; Yang and Kasuya (1994, Table 4, p.626) used Sondhi and Schroeter’s (1987) hybrid time-frequency domain synthesiser which incorporates source-tract interactions and a number of vocal-tract losses, and found differences of up to 7.2% in the formant frequencies of their male subject; Beautemps et al. (1995, Table 2, p.35) used a lossless vocal-tract model with the wall-vibration and lip-radiation taken into account by a correction to the formant frequencies and a correction to the length of the lip-section, respectively, and despite explicit optimisation of their directly-measured area-functions to match the measured and model-synthesised formants, they still obtained errors of up to 10% in the formant frequencies; Story et al. (1996, Table IV, p.549) used a “wave-reflection analog vocal tract model” which includes viscosity, wall-vibration, and lip-radiation losses, in addition to an acoustic side-branch to model the piriform sinuses, and still found deviations from measured formant frequencies up to nearly 15% in $F_1$, up to 16% in $F_2$, and up to nearly 20% in $F_3$.

In row (a) of Table 5.5 we list the rms difference (in Hz) between the measured and synthetic values for each of the first three formant frequencies published in those five studies, and the rms difference in the fourth formant frequencies published by Yang and Kasuya (1994) and by Beautemps et al. (1995). By comparison, row (b) lists the rms differences (in Hz) between the formants synthesised first under model-matched, then (presumed) more realistic conditions (those listed earlier in Tables 5.3 and 5.4, respectively). Both sets of results suggest rms errors on the order of 10% — clearly

<table>
<thead>
<tr>
<th>Formant</th>
<th>Studies</th>
<th>rms Difference (in Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>Fant</td>
<td>5%</td>
</tr>
<tr>
<td>$F_2$</td>
<td>Fant</td>
<td>10%</td>
</tr>
<tr>
<td>$F_3$</td>
<td>Fant</td>
<td>15%</td>
</tr>
<tr>
<td></td>
<td>Baer et al.</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>Yang and Kasuya</td>
<td>7.2%</td>
</tr>
<tr>
<td></td>
<td>Beautemps et al.</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>Story et al.</td>
<td>15%</td>
</tr>
</tbody>
</table>
greater than the 3 – 5% perceptual difference limen suggested by Flanagan (1955). However, it is even more interesting to note that for the first three formant frequencies, the rms errors listed in row (a) of Table 5.5 (those inferred from the studies cited above) are in fact larger than those obtained (and listed in row (b) of Table 5.5) in our evaluation of the discrepancies between the LP and a more realistic vocal-tract model!

That result alone is indicative of the futility of formant correction procedures based on model data. This assertion is indeed confirmed by an experiment in which we attempted to correct the formants listed in Table 5.4, according to the correction charts of Figure 5.13. In particular, we used the cubic-polynomial of Equation 5.21 to correct the formant frequencies, and four separate, linear functions to correct the formant bandwidths. Perhaps not surprisingly, the mean rms error yielded for the re-estimated area-functions (0.33), was found to be only slightly lower than that obtained originally under model-mismatched conditions (0.37). In addition to confirming the futility of model-based formant correction, these results also suggest that the mismatched conditions simulated in the previous section may indeed have been exaggerated, by virtue of our over-reliance on model-generated (or synthetic) formant data. In the next section we therefore consider the inversion problem using real (or measured) formants.

### 5.5.3 Estimation of Area-Functions from Measured Formants

If we are to use our hybrid LP-SM method of inversion to estimate area-functions from real, measured acoustic data, then we are compelled to alleviate the following problems which would otherwise potentially undermine the interpretability of our results. The

<table>
<thead>
<tr>
<th></th>
<th>$F_1$</th>
<th>$F_2$</th>
<th>$F_3$</th>
<th>$F_4$</th>
<th>$B_1$</th>
<th>$B_2$</th>
<th>$B_3$</th>
<th>$B_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>53.5</td>
<td>183.7</td>
<td>251.2</td>
<td>145.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(b)</td>
<td>40.0</td>
<td>160.6</td>
<td>185.5</td>
<td>228.0</td>
<td>40.4</td>
<td>54.5</td>
<td>166.4</td>
<td>173.4</td>
</tr>
</tbody>
</table>

Table 5.5: Root-mean-square (rms) difference (in Hz) in each of the first four formant frequencies and bandwidths, either synthesised or measured, and pertaining to the 33 directly-measured area-functions obtained from the literature. (a): rms difference between measured formant frequencies and those synthesised from the directly-measured area-functions, as published in the five studies referred to in the text (cf. Section 5.5.2.3). (b): rms difference between formants first LP-synthesised with an 8-section representation (cf. Table 5.3), then synthesised using the original section-lengths and areas, with a lossy transmission-line analog (cf. Table 5.4).
first problem (considered in Section 5.5.3.1) concerns the sheer vari ability which is known to plague measurements of formant bandwidths in particular. The second problem (considered in Section 5.5.3.2) concerns the potential mismatch between measured formant bandwidths, and those more appropriate to the LP vocal-tract model. Although both of these problems have long been acknowledged, we still lack a quantitative assessment of, and an accepted method of dealing with, their potentially harmful influences on estimated area-functions.

5.5.3.1 Formant-induced Variability of LP Area-Functions

Whilst formant-tracking has traditionally been concerned with reliably measuring the formant frequencies, we have shown in this chapter that the formant bandwidths do play a vitally supporting role in ensuring uniqueness in the LP-based method of inversion. Indeed, the distinctive relation which we have shown to exist between each formant bandwidth and the corresponding, symmetric component of the LP-derived, logarithmic area-function, is sufficient to suspect that large amounts of inter-frame or inter-repetition variation in bandwidth measurements might induce significant variability in the estimated vocal-tract shapes. As this would clearly be an undesirable input to our forthcoming, articulatory interpretation of the phenomenon of dichotomy, it is of utmost importance to quantify the extent to which measurement variability in either formant frequencies or bandwidths are likely to influence the estimated vocal-tract shapes.

To that end, a preliminary step is taken by computing the relative sensitivity of our shape parameters to variations in the formant frequencies and bandwidths. This is achieved by fitting the distribution of each of the four formant frequencies and the four bandwidths generated (in Section 5.4.2.1) from the 6561 area-functions perturbed about the neutral-tube, using a linear combination of all eight, re-estimated shape-parameters, as shown by the following, multiple-linear regression formulae:

\[
F_n^{(rel)} = \sum_{m=1}^{M/2} \alpha_{n,m} a_{2m-1} + \sum_{m=1}^{M/2} \beta_{n,m} b_{2m-1}, \quad n = 1, \ldots, M/2, \tag{5.22}
\]

\[
B_n^{(rel)} = \sum_{m=1}^{M/2} \alpha_{n,m} a_{2m-1} + \sum_{m=1}^{M/2} \beta_{n,m} b_{2m-1}, \quad n = 1, \ldots, M/2, \tag{5.23}
\]
where the superscript “(rel)” refers to the relative formant parameters, as defined earlier in Section 5.4.2.1. The resulting set of $M = 8$ hyperplanes (each of which passes through the origin), are fully described by the following direction-cosines with respect to each shape parameter:

\[
\alpha_{n,m}^{(F)}, \beta_{n,m}^{(B)}; \quad n = 1, ..., M, \quad m = 1, ..., M / 2.
\]

The coefficients themselves are found such that each hyperplane fits the 6561 data points in a least-mean-squares sense, using a generalised linear least-squares algorithm (Press et al., 1988, p.537) based on singular value decomposition (SVD).

Table 5.6 lists the coefficients of sensitivity thus obtained. In particular, we draw attention to those coefficients along the main diagonal, which quantify the sensitivity in each of the so-called first-order parameter relations, shown earlier to have the most significant coefficients of correlation ($|r| > 0.9$).

Table 5.6: Coefficients of hyperplanar model, obtained from the 6561 pairs of acoustic/vocal-tract shape data generated in Section 5.4.2.1, for each of the first four, relative formant frequencies and bandwidths, in terms of the first eight shape parameters (cf. Equations 5.22 and 5.23). Boldface figures correspond to the so-called first-order parameter relations which were earlier shown (in Table 5.2) to have high coefficients of correlation ($|p| > 0.9$).

<table>
<thead>
<tr>
<th></th>
<th>$a_1$</th>
<th>$a_3$</th>
<th>$a_5$</th>
<th>$a_7$</th>
<th>$b_1$</th>
<th>$b_3$</th>
<th>$b_5$</th>
<th>$b_7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1^{(rel)}$</td>
<td>-0.535</td>
<td>-0.040</td>
<td>-0.040</td>
<td>-0.044</td>
<td>-0.015</td>
<td>-0.017</td>
<td>-0.006</td>
<td>0.018</td>
</tr>
<tr>
<td>$F_2^{(rel)}$</td>
<td>-0.041</td>
<td>-0.516</td>
<td>-0.050</td>
<td>-0.053</td>
<td>-0.022</td>
<td>-0.009</td>
<td>-0.011</td>
<td>0.019</td>
</tr>
<tr>
<td>$F_3^{(rel)}$</td>
<td>-0.048</td>
<td>-0.053</td>
<td>-0.479</td>
<td>-0.071</td>
<td>-0.029</td>
<td>-0.014</td>
<td>-0.007</td>
<td>0.020</td>
</tr>
<tr>
<td>$F_4^{(rel)}$</td>
<td>-0.019</td>
<td>-0.021</td>
<td>-0.026</td>
<td>-0.380</td>
<td>-0.011</td>
<td>-0.006</td>
<td>-0.002</td>
<td>0.007</td>
</tr>
<tr>
<td>$B_1^{(rel)}$</td>
<td>0.228</td>
<td>0.954</td>
<td>0.590</td>
<td>0.197</td>
<td>-0.776</td>
<td>0.032</td>
<td>0.064</td>
<td>0.082</td>
</tr>
<tr>
<td>$B_2^{(rel)}$</td>
<td>-0.164</td>
<td>0.073</td>
<td>0.968</td>
<td>0.255</td>
<td>0.106</td>
<td>-1.867</td>
<td>0.493</td>
<td>0.618</td>
</tr>
<tr>
<td>$B_3^{(rel)}$</td>
<td>-0.340</td>
<td>-0.954</td>
<td>-0.494</td>
<td>0.273</td>
<td>0.089</td>
<td>0.643</td>
<td>-2.184</td>
<td>1.469</td>
</tr>
<tr>
<td>$B_4^{(rel)}$</td>
<td>0.276</td>
<td>-0.073</td>
<td>-1.064</td>
<td>-0.726</td>
<td>0.581</td>
<td>1.191</td>
<td>1.627</td>
<td>-2.170</td>
</tr>
</tbody>
</table>

The analogous coefficients of sensitivity for each of the first four formant bandwidths, are $\beta_{1,1}^{(B)} = -0.776$, $\beta_{2,2}^{(B)} = -1.867$, $\beta_{3,3}^{(B)} = -2.184$, and $\beta_{4,4}^{(B)} = -2.170$, respectively. These
increasingly larger-valued, negative coefficients, were already foreshadowed by the slopes of the relevant nomograms shown earlier in Figure 5.5(b).

In order to obtain a first approximation of the sensitivity of each shape parameter with respect to a unit perturbation in the corresponding, relative formant parameter, we assume that the matrix found above is diagonally dominant, and simply compute the reciprocal of each of those hyperplane coefficients listed along the main diagonal. The inverse-sensitivities thus obtained, are \(-1.87\), \(-1.94\), \(-2.09\), and \(-2.63\), respectively, for each of the first four, asymmetric shape parameters \(a_{2n-1}\), and \(-1.29\), \(-0.54\), \(-0.46\), and \(-0.46\), respectively, for each of the first four, symmetric shape parameters \(b_{2n-1}\). The results of this simple calculation imply that vocal-tract shapes obtained by our hybrid LP-SM method of inversion, are less sensitive to unit variations in the relative bandwidths, than they are to unit variations in the relative formant frequencies. In view of the notorious difficulties of measuring formant bandwidths robustly, this preliminary analysis provides some hope that the bandwidth-induced variations in vocal-tract shapes may not be so disastrous as to obscure important information pertaining, for example, to the phonetic identity of the vowel being considered.

We now turn to the formant data of our four adult, male speakers of Australian English, in order to better assess the detrimental influences of formant variations on estimated vocal-tract shapes. As described in Chapter 3, we used both our phonetic knowledge and the formant-enhancement properties of the NDPS, in order to ensure consistent and relatively noise-free estimates of the formant frequencies in particular. However, the formant bandwidths can be expected to have relatively larger amounts of both inter-frame and inter-repetition variability, which may therefore induce a greater amount of variability in LP-derived vocal-tract shapes, despite the lower sensitivity of those shapes to the bandwidths than to the formant frequencies (as found above). In order to assess the relative influence of formant frequency and bandwidth variations on estimated vocal-tract shapes, the mean and standard-deviation (\( \sigma \)) of each formant frequency and bandwidth were first computed on a per-vowel, per-speaker basis, across the 7 steady-state frames and 5 repetitions. The following two sets of formant data were then prepared for each vowel of each speaker: (i) with the 4 bandwidths fixed at their mean values, the 4 formant frequencies are set equal to \( \pm 1\sigma \) about their respective
mean values, in all possible \((2^4 = 16)\) combinations; and (ii) with the 4 formant frequencies fixed at their mean values, the 4 bandwidths are set equal to \(\pm 1\sigma\) about their respective mean values, in all possible \((2^4 = 16)\) combinations. For each of those two sets of 16 formant-patterns, our hybrid method of inversion was then used to estimate the corresponding groups of 16 area-functions, which in turn were subjected to the alignment procedure described in Section 5.5.1.

Three representative examples of the area-functions thus obtained are shown in Figure 5.14, where the two groups of 16, aligned vocal-tract shapes are superimposed in each graph. In particular, it is fairly consistently observed that the formant frequency-perturbed vocal-tract shapes (the green area-functions) cluster together more closely than the bandwidth-perturbed shapes (the red area-functions). This immediately confirms our suspicions raised in the previous paragraph, that despite the smaller sensitivities of vocal-tract shape parameters to variations in bandwidths, the measured, inter-frame and inter-repetition variations in the bandwidths are sufficiently large (compared with the measured variations in the formant frequencies) as to induce a greater amount of variability in the estimated shapes.

The results are summarised in Figure 5.15, which shows the per-vowel, mean rms dispersion (averaged over the 4 speakers) in the groups of estimated vocal-tract shapes. These results first confirm that the dispersion in vocal-tract shapes caused by \(\pm 1\sigma\) variations in measured formant bandwidths (shown by the plus symbols joined by dashed lines), is indeed consistently larger than the shape-dispersions induced by \(\pm 1\sigma\) variations in the formant frequencies (diamond symbols joined by solid lines). Nevertheless, this quantitative evidence confirms our observations in Figure 5.14, that at least for the formant data at hand, the bandwidth-induced articulatory dispersions are not so significantly larger than the formant frequency-induced dispersions, that they should cause catastrophic distortions of the estimated vocal-tract shapes. Indeed, the mean, formant frequency-induced dispersions for the two back vowels in “hod” and “hoard”, are numerically within the same range as the mean, bandwidth-induced dispersions for some of the front vowels. This finding is all the more remarkable since, as discussed earlier, the measured formant bandwidths were not subject to the rigorous constraints imposed by our methodology for extracting the steady-state formant
Figure 5.14: Illustration of the vocal-tract shape dispersions induced by all combinations of ±1σ perturbations in each of the first four, measured formant frequencies (green area-functions) and bandwidths (red area-functions) of the FC dataset. Top panel: vowel /i/ of speaker C. Middle panel: vowel /u/ of Speaker D. Bottom panel: vowel /a/ of Speaker B. All area-functions are estimated using our hybrid LP-SM method of inversion. As exemplified in these three graphs, it is consistently observed in all 11 vowels of all 4 speakers of the FC dataset, that the bandwidth-induced shape dispersions are larger than the formant-induced shape dispersions.
frequencies; nor did we use a pitch-synchronous acoustic analysis which might have reduced the bandwidth measurement variability caused by the varying position of the analysis window relative to the peaks of glottal excitation in the speech waveform.

In sum, our results indicate that bandwidth-induced dispersions in vocal-tract shape (mean rms error of 0.181) are larger than formant frequency-induced dispersions (mean rms error of 0.097) by approximately a factor of two. It is interesting to note that by comparison, the normalised standard-deviation of the bandwidths themselves (mean value of 0.163) is larger than that of the formant frequencies (mean value of 0.048) by a factor greater than three.\(^6\)

Although studies concerned with the expected measurement variability in formant bandwidths are quite rare, Markel and Gray (1976, p.178) do offer the following

\(^6\) The smaller factor obtained for the dispersions in vocal-tract shapes than for the formants themselves, is partly explained by the coefficients of inverse-sensitivity found earlier.
In light of this and the evidence summarised in Figure 5.15, we must acknowledge the potentially harmful influence of the presumably non-phonetic variations in formant bandwidths, and prescribe a method of dealing with those variations. As the bandwidths do clearly play a role in determining the so-called second-half of shape components which are otherwise missing in the completely lossless model, their phonetic variation (i.e., from vowel to vowel) should perhaps be retained. Towards this end, we have determined to use only the mean of each bandwidth on a per-vowel basis, computed across frames, repetitions, and speakers. This prescription for controlling the non-phonetic variability in the formant bandwidths is all the more pertinent that our vowel classification experiments using simplified cepstra have already shown (in Chapter 4) that the acoustic-phonetic phenomenon of speech-speaker dichotomy is independent of these acoustic parameters.

5.5.3.2 Closed-glottis Correction of Formant Bandwidths

Whilst the sheer variability in measured formant bandwidths can be controlled by using only their per-vowel mean values as recommended above, a serious question remains as to the relevance of those mean bandwidths to the LP vocal-tract model. The potential mismatch stems from the fact that while “true” or measured formant bandwidths reflect all types of losses in the human vocal tract, such as those caused by the glottal source, acoustic viscosity, heat-conduction, wall-vibration, and lip radiation, the LP vocal-tract model has only a single, frequency-independent source of acoustic energy loss located at the glottal end. Hence, as stated by Fant (1980, p.85):

“The bandwidths we need for the inverse LPC-based transforms are the bandwidths of a production model which has losses at the glottis only and lacks the cavity wall shunt.”

However, as we have earlier argued (in Section 5.5.2.3), model-based formant correction procedures are not easily formulated, nor do they appear to be particularly effective. In this vein, Fant’s (1980, p.85) very next, more general statement is perhaps a more appropriate guide to pursuing the problem:
“From the true formant frequencies and bandwidths we thus have to make a best guess of what bandwidths the LPC model would generate.”

If model-based data are to be avoided, then we are naturally drawn towards measured bandwidth data, in search of a clue as to the necessary “best guess”.

In that regard, it is relevant to note that any measured bandwidth is known to be approximately an additive combination of its individual components (in Hz), which themselves are determined by the configuration-dependent contribution from each of the separate sources of loss in the vocal-tract (e.g., Fant, 1960; Flanagan, 1972). Bandwidths measured from the acoustic speech signal are thus the total sum of their individual components arising from glottal, vocal-tract internal, lip-radiation and other losses; by contrast, the LP vocal-tract model “would generate” bandwidths from only the glottal resistance. The question then arises whether one can approximately compensate for the losses which are missing in the LP model, by subtracting those components of each bandwidth which might have been measured had there been no source of loss at the speaker’s glottis.

Such measurements of so-called “closed-glottis bandwidths” have indeed been made on human speakers (Fant, 1962; Fujimura and Lindqvist, 1971), by applying a sweep-tone excitation signal transcutaneously at the throat while the speaker maintains a static articulatory configuration with closed glottis, and by measuring the acoustic response near the speaker’s lips. Using those empirical data, Fant (1972) then brought acoustic theory to bear on the problem of deriving a formula for each of the first three bandwidths, in terms of the first three formant frequencies and an average, fourth formant frequency. More recently, Hawks and Miller (1995) used those empirical data to derive a more general equation which is able to predict the bandwidth of any of the formant frequencies up to nearly 5kHz, given only the centre-frequency of that formant.

If we can assume that the per-vowel, mean values of the measured bandwidths embody all of the natural sources of loss which arise for the given vocalic configuration, and that Hawks and Miller’s (1995) equation embodies the relation between frequencies and bandwidths of closed-glottis formants, then we might expect that by subtracting thus estimated closed-glottis bandwidths from the measured bandwidths, we would obtain an approximation of the per-vowel, mean glottis-only bandwidths. It is indeed
the glottis-only bandwidths which the LP vocal-tract model ideally would require in order to yield area-functions which (presumably) more closely resemble those actually produced by a given speaker. Clearly, it is difficult to test this particular hypothesis without simultaneously-measured acoustic and vocal-tract shape data; however, we may at least test the viability of the procedure by determining whether the subtraction would yield admissible (i.e., positive and non-zero) bandwidths.

To that end, in Figure 5.16 is shown Hawks and Miller’s (1995) bandwidth-versus-frequency curve (assuming their default conditions for an adult male speaker). Diamond symbols: the per-vowel mean of each of the first four formants computed over the 7 steady-state frames, 5 repetitions, and 4 adult, male speakers of the FC dataset.

Figures 5.16: Illustration of the viability of closed-glottis bandwidth correction. Solid curve: Hawks and Miller’s (1995) empirically-derived equation relating the closed-glottis bandwidth of a formant with its centre-frequency (assuming their default conditions for an adult male speaker). Diamond symbols: the per-vowel mean of each of the first four formants computed over the 7 steady-state frames, 5 repetitions, and 4 adult, male speakers of the FC dataset.

the glottis-only bandwidths which the LP vocal-tract model ideally would require in order to yield area-functions which (presumably) more closely resemble those actually produced by a given speaker. Clearly, it is difficult to test this particular hypothesis without simultaneously-measured acoustic and vocal-tract shape data; however, we may at least test the viability of the procedure by determining whether the subtraction would yield admissible (i.e., positive and non-zero) bandwidths.

To that end, in Figure 5.16 is shown Hawks and Miller’s (1995) bandwidth-versus-frequency curve (assuming their default conditions for adult male speakers), together with the mean frequencies and bandwidths of each of the first four formants of the 11 vowels of our adult, male speakers of the FC dataset (diamond symbols). This Figure shows that our measured bandwidth data do follow the expected trend of higher values for higher formant frequencies. More importantly, it clearly shows that our mean data do consistently lie above the curve, and that the difference (or the vertical distance) between each mean bandwidth (or data point) and the corresponding, closed-glottis bandwidth (along the curve), would indeed yield a positive and non-zero, “glottis-only”
Whilst the corrected bandwidths (listed in Appendix G, together with the measured, mean formant frequencies and bandwidths for each vowel) do confirm the viability of the procedure, it would also be interesting to know how, and to what extent, these bandwidth corrections influence the estimated vocal-tract shapes themselves. Shown in each of the 11 panels in Appendix G, are superimposed the aligned area-functions estimated from our mean formant data before (dashed line) and after (solid line) bandwidth correction. Those graphs show that the consequences of our proposed method of bandwidth correction are not only harmless, but may even be considered beneficial on the whole. Differences between the pairs of vocal-tract shapes appear to be the largest for the quasi-neutral vowel /ə/ (which is also somewhat shorter after bandwidth correction), the fronted /ʉː/ (which develops a more well-defined constriction), and the back vowel /o/ (whose front cavity expansion reduces in volume to a somewhat more realistic level). It is also interesting to note that after bandwidth correction, the two, high-back vowels /ɔ/ and /u/ are no longer found to require the addition of a fifth, synthetic formant in order for the inversion method to determine an optimum vocal-tract length (cf. Section 5.4.1).

Our qualitative observations are quantitatively confirmed in terms of the rms differences between the pairs of (logarithmic) area-functions, shown by the square symbols (joined by dotted lines) in Figure 5.15 (see the end of the previous section). That graph clearly indicates that the shape-related consequences of our proposed method of bandwidth correction, are less severe than either the formant bandwidth- or frequency-induced dispersions found in the previous section. Indeed, the overall mean rms error (0.03) is even less than that found earlier for the 33 directly-measured area-functions re-estimated under model-matched conditions (0.04, see Section 5.5.2.1).

It is highly encouraging that the consequences of formant bandwidth correction do not appear to be as drastic as the literature would generally seem to suggest. However, we must concede that our proposed method of determining the mean, “glottis-only”

---

It is important to note that such, “glottis-only” bandwidths would represent only an average of the quasi-periodic bandwidth values, owing to our speech analysis frame-lengths typically extending over more than one glottal cycle.
bandwidths which ought to be better matched with the LP vocal-tract model, must be more conclusively put to the test, perhaps with the help of simultaneously-measured acoustic and vocal-tract shape data. Until such time, we may rest assured that at least for the formant data at hand (the FC dataset), the proposed correction applied to the per-vowel mean bandwidths, should not endanger the phonetic interpretability of the estimated area-functions, nor their articulatory manifestations of the speech-speaker dichotomy, which we wish to investigate in the forthcoming chapter.

5.6 Concluding Summary

Our aim in this chapter was to find a method of mapping the geometry of the human vocal-tract from the acoustic speech signal, in order to facilitate an extension of our acoustic-phonetic investigations of the speech-speaker dichotomy (in Chapter 4) into the domain of speech production. As outlined in Section 5.1, we sought a method of inversion which would satisfy the following two criteria: uniqueness and formant-based parameterisation. These criteria served as guideposts in our bipartite rationale (in Section 5.2), where we reviewed the strengths and the limitations of the Schroeder-Mermelstein (SM) model and the linear-prediction (LP) vocal-tract model.

The main strength of the SM model is its description of vocal-tract area-functions in terms of orthogonal parameters which relate quasi-linearly and uniquely to the resonance frequencies of the acoustic-tube. Its main limitation in the context of area-function estimation, is that only one-half of the required information is available in the acoustic speech signal (i.e., only the formant frequencies). Hence, the estimated vocal-tract shapes are ambiguous in their symmetric components, unless further constraints are introduced.

The LP model, on the other hand, is well-known to yield a unique vocal-tract shape (for a given vocal-tract length), and is notable for its mathematical tractability and low computational complexity in comparison with other inversion methods which have been proposed. However, while the formant bandwidths are implicated as the necessary second-half of acoustic information required to obtain a unique LP area-function, it has never been explained how they contribute to the shape-related uniqueness of those area-functions. Furthermore, the LP-based method of inversion suffers from a number of
limitations, including the analysis artefact of an upper limit $L_{\text{max}}$ on allowed vocal-tract length, and the coarse, step-wise representation which inhibits comparison of area-functions of different length.

With the aim of elucidating the uniqueness property of the LP inversion method, and perhaps overcoming some of its inherent limitations, we then used the SM model as a probe to examine the formant-dependence of LP-derived vocal-tract shapes. In Section 5.3 we first validated the SM model within the LP modelling framework, and thereby showed that those two models are indeed isomorphic in regard to the quasi-linear relation between the formant frequencies and the antisymmetric (cosinusoidal) perturbations of the uniform (logarithmic) area-function. We then derived a theoretical proof of the dependence of the mean formant bandwidth on the glottal reflection coefficient, and thereby provided the motivation for regarding the relative bandwidths more pertinent to shape components of LP-derived area-functions. Remarkably, our empirical investigations then revealed those bandwidth-related shape components to be the symmetric (sinusoidal) perturbations of the uniform (logarithmic) area-function.

Given the acknowledged (but to date largely unexplained) importance of formant bandwidths in securing uniqueness in LP-derived area-functions, our proposed method of vocal-tract shape parameterisation (given in Section 5.4) is therefore an important contribution to our understanding of the uniqueness property of the LP inverse method. More importantly, it overcomes the inherent discreteness of LP-derived vocal-tract shapes, and provides a smooth representation by way of a compact and orthogonal set of parameters which also lend themselves to an acoustic-phonetic interpretation. Furthermore, the artificial limitation of $L_{\text{max}}$ is easily overcome by invoking the very property of the SM model which explicitly relates higher-resolution shape components with the higher formants. In particular, the vocal-tract shape is estimated to a specified degree of spatial resolution, while using a sufficient number of vocal-tract sections to ensure that the highest available formant still lies within the spectral range of the vocal-tract model.

In Section 5.4.2 we first used our hybrid LP-SM method of inversion to evaluate our proposed method of area-function parameterisation. In particular, we obtained high coefficients of correlation between each of the formant frequencies and bandwidths
synthesised from 6561 different area-functions, and the corresponding shape-parameters re-estimated from those formant data. The acoustic relevance of our shape parameters thus confirmed, their physical relevance was assessed by representing 33 directly measured, human vocal-tract area-functions obtained from the literature. On average, the most important shape-components of those area-functions were found to be the first two pairs of both antisymmetric and symmetric components. This evidence confirmed the phonetic importance of the first two formant frequencies of vowels; however, it also confirmed the importance of the bandwidths in describing unique vocal-tract shapes.

The hybrid method of inversion itself was finally evaluated in Section 5.5, with the help of a new method for aligning and quantifying the difference between similarly-shaped area-functions. We first re-estimated the 33 directly measured area-functions obtained from the literature, using the first four formant frequencies and bandwidths synthesised under model-matched conditions (i.e., a vocal-tract model with a glottal resistance as the only source of loss, and the number of vocal-tract sections equal to twice the number of formants). Whilst the phonetic range of variation in vocal-tract length was found to be exaggerated upon re-estimation, the vocal-tract shapes were fairly accurately re-estimated, with a mean rms error of 0.12. More realistic conditions were then imposed, by re-estimating the area-functions from formants synthesised using a more lossy vocal-tract model, and by retaining the original section-lengths and areas. The mean rms error between original and re-estimated shapes under model-mismatched conditions, was found to be 0.37.

In the hope of finding a reasonably accurate method of compensating for the LP model’s simplified assumptions in regard to vocal-tract losses and number of sections, we presented a more comprehensive version of the “formant frequency conversion chart” first proposed by Wakita (1979). Whilst we were able to fit the model-generated formant data with a third-order polynomial function in logarithmic frequencies, the analogous conversion chart for formant bandwidths did not suggest the possibility of finding a single functional description. Indeed, the futility of approaching the formant correction problem from a purely model-based point of view, was further reinforced by our finding that on average, errors between formant frequencies measured from the speech waveform and those re-synthesised from the directly-measured area-functions
are in fact larger than the corresponding errors between the formant frequencies synthesised, respectively, under model-matched and model-mismatched conditions.

In Section 5.5.3 we then turned to our measured vowel formant data (of the FC dataset). As a prelude to examining formant-induced variability using those real data, we first established that LP-derived vocal-tract shapes are actually less sensitive to unit perturbations in the relative formant bandwidths, than to unit perturbations in the relative formant frequencies. Despite this lower sensitivity, the inter-frame and inter-repetition variability in our measured bandwidth data was large enough to induce dispersions in estimated vocal-tract shapes, approximately twice that induced by variability in the formant frequencies. In view of our forthcoming, articulatory investigation of the speech-speaker dichotomy which we had earlier explained in terms of the formant frequencies (in Chapter 4), our prescription was therefore to control the inter-frame, inter-repetition, and the inter-speaker variations in each formant bandwidth (while retaining their phonetic variation, the importance of which was indirectly suggested by our results in Section 5.4.2.2), by reducing the bandwidths to their mean values computed on a per-vowel basis.

Whilst the measurement variability in formant bandwidths was thus controlled, their relevance to the LP vocal-tract model was still questionable. Our approach to dealing with this problem was then to recognise that the bandwidths required by the LP model are those which would arise from a vocal-tract having only a resistive glottal termination. We therefore used Hawks and Miller’s (1995) empirically-derived equation to determine the so-called “closed-glottis” bandwidths of our mean formant data, then subtracted those from the original values in order to approximate the required, glottis-only bandwidths. Although a more conclusive proof of this procedure would perhaps require simultaneously-measured acoustic and area-function data, our results do support its viability, and furthermore indicate only minor differences in the vocal-tract shapes re-estimated after bandwidth correction.

Finally, the detailed evaluation presented in Section 5.5 raises our confidence in LP-derived and parameterised vocal-tract shapes obtained by the hybrid LP-SM method described in Section 5.4, which itself secures both uniqueness and resonance-based parameterisation of estimated vocal-tract area-functions. In the following chapter we
therefore proceed to use our method of inversion and area-function parameterisation, with a view toward an investigation of the speech-speaker dichotomy in the speech production domain.