VOCAL-TRACT AREA-FUNCTION PARAMETERS FROM FORMANT FREQUENCIES

Qiguang LIN and Gunnar FANT

Dept. of Speech Communication and Music Acoustics
Royal Inst. of Technology, 10044 Stockholm, Sweden

ABSTRACT

The basic three-parameter vocal tract area function model of Fant [2] has been extended to allow for asymmetry and variable length of the tongue hump section as well as variations of the VT overall length.

An important class of articulations are those that conform with this model, mostly vowel sounds and consonants without apical modification or a secondary posterior articulation. Staring out from a theory of small perturbations [1,4], we calculate the shift in the VT parameter accompanying a certain set of shifts in F1, F2, and F3. This is done from a set of three linear differential equations expressing the shifts in each of F1, F2, and F3 as the sum of the contributions from the unknown shifts in each of the VT parameters. These are solved at each small step of a pathway from a reference starting point to the target formant-pattern. The overall length is optimized by reference to F4 and some available measured data.

A resynthesis of a time varying F-pattern from a set of articulatory targets has been attempted with a possible application in articulatory synthesis.

1. INTRODUCTION

The purpose of this paper is to infer VT configurations from formant frequencies for vowel-like sounds. It is well known that without additional appropriate constraints there exist an infinite number of solutions and that many of them are completely unrealistic. In the present study, the constraint imposed is that a VT configuration should be generated by means of a parametric VT model. For this purpose, the three-parameter VT model of Fant [2] is extended to be more feasible to deal with and capture more essential features of the production of vowels.

In addition to the overall tract length, $l_{an}$, there are still three main independent parameters in the present version. Altogether, they specify the size ($A_c$) and location ($X_c$) of the VT constriction and the state of the lip section ($l_d/A_o$). Once these parameters are known, the underlying VT profile, in terms of area function is defined.

To derive these articulatory parameters, a set of three linear differential equations are formulated. The basic algorithm is to first determine the amount of variation in formant frequency due to a (small) known perturbation of the articulatory parameters (namely, the sensitivity function). This knowledge is then made use of to determine the variation in these parameters for a desired shift of the formant frequency. It is thus an iterative procedure. For the time being, we use the first three formant frequencies to solve for the three model parameters, $A_c$, $X_c$, and $l_d/A_o$. The overall length of the VT $l_{an}$ is assumed to be known in advance and is optimized with reference to $F_4$.

The relation between the acoustic and articulatory data is non-linear. If the step of formant frequency during one iteration is too large, then this linear algorithm may fail to converge. However, the gross linearity of the data between the acoustic and articulatory domains can be preserved by dividing that step into a number of smaller sub-steps [1]. Moreover, the success of the algorithm depends highly upon an appropriate choice of the starting position for the iterative calculation. A proper selection of this starting position can not only minimize the distance between the initial point and the target (given) formant-pattern thereby reducing computation time, but also reduces the risks that searching process falls in the so called "forbidden area" [1]. At present, three rules are implemented to assign a starting position based on the given information of $F_1$ and $F_2$.

The inverse transformation approach has been tested on a set of Swedish vowels [3]. Reasonable VT area functions have been inferred. We have also attempted to derive the VT area configuration for the Russian vowel [a] whose area function is known [2]. The result indicates that an asymmetric tongue-hump outline is better suited for this vowel than a symmetric one, in view of both the resultants $F_4$ and area function.

The VT parametric model has also been applied to an articulatory synthesis scheme. There are two ways of deriving the time varying VT area function. One is directly from analysis of the F-pattern sampled at suitable instants followed by a recovery of VT area functions from the inverse transformation. The second method, which is oriented towards the development of rule systems, is to interpolate $A_c$, $X_c$, $l_d/A_o$, and $l_{an}$ on a basis of prescribed articulatory targets (anchoring points), by resorting to some weighting functions. It is found that such interpolation renders a better result if an intermediate state is included. Synthesis of dynamic pattern of diphthong-like syllables such as [a] has been attempted in these two ways.

2. THE VT MODEL

The curvature of the tongue hump need not be one and the same for all of vocalic sounds. Thus, the first extension of Fant’s VT model [2] is to allow for symmetry and variable length of the tongue hump [7]. The second modification is the utilization of circular functions instead of hyperbolic ones, for the specification of the horn-shaped constriction. This is to gain computation efficiency.

The area function for the horn anterior and posterior to the constriction centre, $A(x)$, is expressed in one of the two alternative forms:

$$A(x) = A_c + (8 - A_c)[1 - \cos(0.1 R_f \pi x/X_f)]$$

$$A(x) = A_c + (8 - A_c)[1 - \cos(0.1 R_f \pi x/X_f)^2]$$

where $R_f$ and $X_f$ control the flaring rate of the area function from the constriction centre $X_c$. $X_f$ also determines the extension of the hump region. Only a portion of the cosine waveform is used thereby avoiding an oscillatory profile and allowing for an adjustable length of the hump section. The cross-sectional area beyond the hump region is set to $0 \text{ cm}^2$.

The displacement from $X_c$ is denoted $x$. For the horn section posterior to the constriction centre, $R_f$ and $X_f$ are replaced by $R_h$ and $X_h$, respectively. Note that for a given $A_c$, $R_f$ and $X_f$ eq. (1b) delivers a longer effective constriction length than eq. (1a) does. It is found that for the back vowels eq. (1a) is more
suited, while for the front vowels, eq. (1b) is more applicable.

It should be pointed out that the laryngeal tube is not directly specified in the VT model. An ad hoc choice of the laryngeal configuration must be made beforehand.

Finally, a number of subsidiary parameters have been introduced. They are used with some or all of the basic model parameters, enabling a dynamic control of e.g., the overall VT length and/or $R_1$ and $R_2$. But these will be treated elsewhere [6].

An example of the area function generated by the present model is provided in Fig. 1. Note that we have chosen the lower incisors as the origin of $X_e$. The length between this origin to the glottal end is denoted as $l_{cog}$. Hence, the overall length of the VT is $l_{tot} = l_{cog} + l_c$. The notation $l_c/A_0$ specifies the lip protrusion by $l_c$ and the lip opening $A_0$.

![Fig. 1. An example of VT configuration generated by the model.](image)

3. ALGORITHM OF INVERSE TRANSFORMATION

The algorithm we have adopted is based on a perturbation analysis of the differential contribution of each VT model parameters to each of $F_t$, $F_z$, or $l_{tot}$, perturbed with a small amount, and the sensitivity function $\delta F_t/\delta P_i$ are calculated. Inserting $\delta F_t$ and $\delta F_z$ into eq. (2) we arrive at an updated VT configuration. The formant frequencies corresponding to this new configuration are evaluated and compared with the goal of that step of calculation. If the error is small we proceed to the next segment (or stop if it is the last segment), otherwise the loop is run through again in the present segment until a satisfactory result is yielded. The algorithm will converge providing that the specific formant-pattern at each step does not drop in the "forbidden area" [1]. An error threshold of 2% is set for intermediate stages and 1% for the final stage. Hereby, the computation burden is relaxed while the final accuracy is still preserved. An example of parameter convergence from a starting point to a target is shown Fig. 2.

![Fig. 2. Parameter convergence from a starting position (position 0) to a target (position 7).](image)

4. SOME RESULTS OF INVERSE TRANSFORMATION

The described system is used to infer the VT profiles for ten Swedish vowels based on the measured formant frequencies in [3]. The result is presented in Fig. 3 (see also [5]). Two clusters of vowels are observed, at $X_e = 10.5$ cm, [a], [i], [y], and at $X_e = 10.5$ cm, [u], [a], [o], respectively. There is an empty zone, $X_e > 1200$ cm, $X_e < 1200$ cm. However, it should be pointed out that the vowels in Fig. 3 pertain to long Swedish vowels. With short vowels included, parts of the empty zone may be filled in. The vowel [ae] appears not unexpectedly at an extreme $X_e$ close to the glottal termination.

The calculated $A_c$ of these vowels also conforms well with expectation, especially with respect to the relation across vowels. For instance, [i] and [u] have the smallest $A_c$ value, [y], [a], [a], and [o] have slightly greater $A_c$. The largest values of $A_c$ are as expected found with [e], [o], and [ae].

![Fig. 3. The resultant model parameters of the Swedish vowels (from [5]).](image)
The lip parameter \( I_p/A_o \) for the ten vowels is given in Tab. 1. In each column, the lip parameter becomes more spread from the top row toward the bottom. The left column pertains to front vowels, \( X_C < 6 \) cm and the right column to "back" vowels, \( X_C > 10 \) cm. These values are realistic and confirm the maximal rounding of [u] and [e].

Tab. 1. Calculated lip parameter \( I_p/A_o \) of the Swedish vowels

<table>
<thead>
<tr>
<th>Vowel</th>
<th>( I_p/A_o )</th>
<th>[u]</th>
<th>[a]</th>
<th>[o]</th>
<th>[e]</th>
<th>[æ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[u]</td>
<td>1.00/8</td>
<td>[u]</td>
<td>1.00/2</td>
<td>[u]</td>
<td>1.00/2</td>
<td>[u]</td>
</tr>
<tr>
<td>[y]</td>
<td>1.10/15</td>
<td>[a]</td>
<td>1.20/5</td>
<td>[æ]</td>
<td>0.80/8</td>
<td>[æ]</td>
</tr>
<tr>
<td>[i]</td>
<td>1.20/0</td>
<td>[a]</td>
<td>1.25/5</td>
<td>[æ]</td>
<td>0.80/8</td>
<td>[æ]</td>
</tr>
<tr>
<td>[e]</td>
<td>0.24/8</td>
<td>[æ]</td>
<td>0.85/8</td>
<td>[æ]</td>
<td>0.80/8</td>
<td>[æ]</td>
</tr>
</tbody>
</table>

In Fig. 3, additional calculations have been undertaken for [e] and [o]. These demonstrate aspects of compensatory articulation. The constriction of the tongue hump of both vowels is located in the vicinity of F3-F2 proximity (cf. references [2, 5]). For this category of articulation, it is crucial to select an appropriate starting point. Otherwise the algorithm will converge on the wrong side of the proximity although the difference is not a direct input to the algorithm, the calculated F4 is close to the measured F4. Dynamic patterns of VT configurations derived this way can add to general knowledge needed for articulatory-oriented synthesis.

![Fig. 5. Inverse transform applied to the syllable [ja]. A comparison between the resultant formant pattern and the measured data (from [5]).](image)

![Fig. 6. Resynthesis of [ja] from the articulatory anchoring points. The resultant formant pattern.](image)

5. RESYNTHESIS FROM ARTICULATORY ANCHORING POINTS

We attempt to derive dynamic patterns of VT configuration by linearly interpolating the model parameters, \( A_o, X_C, I_p/A_o, \) and \( \text{length} \), at desired instants, based on given articulatory anchoring points. Squared sinusoidal functions (first 1/4 period) have been adopted as the weighting functions for the interpolation which have a relatively stable interval both at the onset and offset.

One problem is the choice of trajectory of the
articulators. As an illustration, consider again the syllable [ja]. A simple linear sequencing of states from [j] to [a] seems unnatural, for one would have to maintain the relatively small tongue constriction area $\Delta r$, typical for [j] and [a], throughout the whole course from the front articulation to the back. It seems more probable that during the movement a neutral state is involved, though this state is not completely reached. So, for [ja], we have three articulatory anchoring points: [j], [i], and [a], implicating that three weighting functions are needed. It is important to note that the sum of these weighting functions at any instants should be equal to 1.

The effect of the intermediate pathway through [i] on the transition from [j] to [a] can be studied in Fig. 6. A fair match to the original in Fig. 5 is achieved when the neutral state is included.

6. CONCLUSIONS

An extended version of Fant’s VT model [2] is described by introducing greater flexibility. The model is defined in the area function domain and can be used for the study of vowel-like segments. This new version sets the constraints in the inference of VT configurations from formant frequencies. The basic algorithm of the inverse transformation is based a perturbation analysis on the differential contribution of each model parameter to each of the first three formants. The procedure to iteratively search for the target from the selected starting position is described. The algorithm is used to derive the VT area function for the ten Swedish vowels. It is found that reasonable results have been achieved. The area function derived by the algorithm for the Russian [a] is very similar to the real area function given in [2], especially, when an asymmetry of the tongue hump is adopted. The preliminary results of resynthesisizing dynamic patterns of the VT configuration from articulatory anchoring points reveal the need for a neutral state transitional target. A fair match between formant-patterns so obtained in synthesis and the measured data is achieved.

Some difficulties, concerning the selection of initial position of the iterative calculation, the compensatory articulation around the region of $F_3$-$F_2$ maximal proximity, setting of appropriate size for the laryngeal tube, and the choice of the length $k_{bop}$ have been pointed out. Further development will be to solve these difficulties. We also intend to gain more experience in the resynthesis from some articulatory anchoring points, especially the effect of the intermediate pathway.

REFERENCES