TRYING TO DETERMINE PLACE OF ARTICULATION OF PLOSIVES WITH A VOCAL TRACT MODEL

ALAIN SOQUET, MARCO SAERENS, AND PAUL JOSPA

Institut de Phonétique and IRIDIA
Université Libre de Bruxelles – CP 110
50, av. Franklin Roosevelt 1050 Bruxelles – Belgium

Abstract

We have recently used the Distinctive Regions and Modes theory ([12]), coupled with a neural controller ([15]), to produce an acoustic-articulatory inversion of a vocal tract model ([17]). This paper presents results on the possible detection of the place of articulation of plosives on the basis of this inversion scheme.

1 Introduction

Mrayati, Carré & Guérin ([12]; see also [3]) have recently presented a theory of speech production based on distinctive modes and on distinctive spatial regions along the vocal tract. This theory provides a framework for articulatory speech synthesis ([13]). It supplies relationships between the variations of the first three formants and the cross sectional areas of eight vocal tract regions of the model. Previous work has shown that such a-priori qualitative knowledge can be used to control and invert non-linear physical processes with a neural network ([15]). In this work, the relationships between the cross sectional areas of the regions and the formant variations are used to provide an acoustic-articulatory inversion of a vocal tract. Acoustic-articulatory inversion is a one-to-many nonlinear problem. It is usually managed by generating articulatory vectors in the articulatory space, and computing the corresponding acoustic parameters. Then, a look-up table can be constructed, providing the relationships between acoustic parameters and articulatory vectors ([11], [1], [7]).

A previous paper ([17]) has shown that a network is able to learn to invert the process, for the eleven French oral vowels. The addition of a constraint on the average volume of the vocal tract allows the system to provide more realistic vocal tract shapes, and clearly improves the convergence rate of the network. These results have been extended to a 30-sections vocal tract by introducing a continuity constraint, and the inversion has been generalized to the vowel space ([18]).

In this paper, the inversion scheme is used to provide an articulatory gesture in the neighbourhood of plosives. This gesture is then analysed to locate the candidates for place of articulation.

Bailly et al. [2] are currently studying a similar, but more ambitious problem: They use Jordan's approach ([5]) to control Maeda's articulatory model ([9]).

2 The Vocal Tract Model

Vocal tract shapes are generated in the framework of the so-called Distinctive Regions and Modes theory ([12], [3]). The model involves an acoustical tube closed at one end (glottis), and open at the other (lips) (Figure 1). This model is based on the study of acoustical properties of vocal tract shapes, compared to those of a neutral uniform tube. For the three formants model, eight regions of different length (the distinctive regions) can be defined. Varying the mean cross sectional area of each of these regions induces specific and quasi monotonic formant variations. The eight regions will be denoted as -A, -B, -C, -D, and D, C, B, A.

Figure 1: Vocal tract divided in 30 parts and 8 different regions.

---

1 The following text presents research results of the Belgian National incentive-program for fundamental research in artificial intelligence initiated by the Belgian State, Prime Minister's Office, Science Policy Programming. We would like to thank René Carré for helpful discussions, and Philip Miller for his help in the draft of the English text.
The vocal tract is divided into thirty parts of equal length. Each part belongs to one region and has the qualitative behaviour of this region (See Figure 1). The cross sectional areas for the first region (A) are scaled from 0.8 cm² to 3.0 cm², and the remaining ones from 0.5 cm² to 15.0 cm². The effective length of the acoustic tube has been set to 19 cm.

3 The Neural Controller

A neural network is used to provide the cross sectional areas to the vocal tract model, when the first three target formants are given as input (Figure 2). Standard back-propagation cannot be used directly for the controller because the optimal control parameters are not known.

Therefore, we use a specialized learning scheme based on an approximation of the backpropagated error that allows adaptive control with the neural network ([16]).

![Figure 2: Architecture of the system.](image)

To perform the inversion (see Figure 2), the following three steps are iterated until the vocal tract model produces the target formants:

1. The neural network is given target formant values.
2. The outputs of the network supply the vocal tract cross sectional areas, and the corresponding formant values are computed (we use an algorithm developed by [8]).
3. The difference between these values and the desired formants are used to correct the connection weights of the network with the modified version of back-propagation algorithm.

By training, the neural network learns to supply the correct cross sectional areas for the production of the target formant values.

The controller is a network with three layers (one hidden layer). Every unit of each layer is connected with the units of the adjacent layers. There are three input units (corresponding to the first three formant values), ten hidden units, and thirty output units (corresponding to the thirty parts of the vocal tract).

The error used for the back-propagation algorithm in the neural network is composed of three terms: The difference between the actual and the target formants, a constraint on the average volume of the vocal tract, and a continuity constraint:

\[
E = \sum_y \left[ \sum_{i=1}^{3} (F_i^y - F_i^y_d)^2 \right] + k_1 \left[ \left( \sum_{i=1}^{30} L A_i^y \right) - V_0 \right]^2 \quad (1)
+ k_2 \sum_{i=1}^{29} (A_i^y - A_{i+1}^y)^2
\]

where the \(F_i^y\) are the formant values computed through the transfer function of the tube ([8]), the \(F_i^y_d\) are the target formants, \(L\) is the length of a part, the \(A_i^y\) are the corresponding areas supplied by the network, \(k_1 = 5 \times 10^{-5}\), \(k_2 = 2 \times 10^{-3}\), and the average volume \(V_0 = 85 \text{ cm}^3\).

This way, the network approximates the nonlinear mapping from the acoustic parameters (the three first formants) to the articulatory space (the cross sectional areas). The net provides one possible solution to this problem and, since it is a one-to-many problem, constraints are introduced in order to reduce the number of possible solutions. Hence, we observe that the different mapping obtained with different initial weights are quite similar.

4 Experiment

The network is first trained on the 11 French oral vowels (we use values published by [10]), then, the training set is generalized to the whole vowel space (see [18]).

After this training, the network approximates the nonlinear mapping from the acoustic parameters (the three first formant values) to the articulatory space (the cross sectional areas). The net is used to provide vocal tract shapes in the neighbourhood of consonant plosives. These shapes are then used to locate a possible constriction place. This allows us to establish whether there is a correlation between this constriction place and
the real place of articulation of the plosive. Indeed, Mayzat, Carré & Guérin (144) claim that the different regions of the acoustic tube correspond to precise places in the vocal tract. For instance, labials are associated with the region A, dentals with the region B, and palato-velars with the regions C and D.

It is well known that important cues for identification of place of articulation of plosives are located in the formant transitions ([4]) and burst spectrum ([19]). In this work, we only take into account the formant values to realize the inversion.

Speakers. Two Belgian male subjects, native speakers of the French spoken in the Brussels area, and with university education were employed.

Recording procedure. The VCV items were recorded with a Studer A310 tape recorder in an anechoic room through a Neumann U88 microphone. They were sampled at 20 kHz with the Maccspeech Lab software on a Macintosh II.

Items. The speakers were asked to produce VCV items, C being one of the six plosives [p, t, k, b, d, g] and V one of the five vowels [a, e, i, y, u]. There were 5 x 6 x 5 = 150 items for each speaker. The sequence consisted of three blocks of 50 items in random order.

Acoustic analysis. The formant values were manually extracted with the Maccspeech Lab software, at two different locations, for both adjacent vowels: The middle of the stable portion of the vowel (t2) and the end of the vocalic transition (t3). We were unable to extract these values for 4 items, the transitions being not detectable. The formants are provided as input to the network, which associates vocal tract configurations. Two different cues are computed on the vocal tube, a static cue, which simply corresponds to the sections at t2, and a dynamic cue, which is:

\[ I_i(t_e) = \frac{A_i(t_2) - A_i(t_e)}{A_i(t_2) - A_i \text{ min}} \]  \hspace{1cm} (2)

The place of articulation of the plosive is determined on the basis of the constriction deduced from the two different cues and the two adjacent vowels. The final decision is taken by a vote of the different knowledge sources.

Confusion matrix is presented in Table 1. We obtain 72.0% identification of classified places and 21.6% of ambiguous cases. Table 2 shows that the dynamic cue is more reliable than the static one, but provides more ambiguity.

There is indeed a correlation between the place of articulation of the plosive and the constriction of the tube. However, a more detailed analysis of the results shows a great influence of the context on the behaviour of the tube. This is not surpring provided that cues for place of articulation are known to be context-sensitive ([120]; [6]). For instance, for context [i], the constriction of the vowel is palato-velar, and remains during the transition. In this particular case, the dynamic cue is much more reliable.

<table>
<thead>
<tr>
<th>Table 1: Total results for the 296 VCV items.</th>
<th>produced</th>
<th>labial</th>
<th>dental</th>
<th>velar</th>
</tr>
</thead>
<tbody>
<tr>
<td>identified</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>labial</td>
<td>75</td>
<td>13</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>dental</td>
<td>5</td>
<td>47</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>velar</td>
<td>6</td>
<td>17</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>ambiguous</td>
<td>12</td>
<td>23</td>
<td>29</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Results for static cue (upper table) and dynamic cue (lower table).</th>
<th>produced</th>
<th>labial</th>
<th>dental</th>
<th>velar</th>
</tr>
</thead>
<tbody>
<tr>
<td>identified</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>labial</td>
<td>57</td>
<td>10</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>dental</td>
<td>1</td>
<td>27</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>velar</td>
<td>6</td>
<td>15</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>ambiguous</td>
<td>34</td>
<td>48</td>
<td>38</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>produced</th>
<th>labial</th>
<th>dental</th>
<th>velar</th>
</tr>
</thead>
<tbody>
<tr>
<td>identified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>labial</td>
<td>51</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>dental</td>
<td>4</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>velar</td>
<td>3</td>
<td>8</td>
<td>35</td>
</tr>
<tr>
<td>ambiguous</td>
<td>40</td>
<td>58</td>
<td>51</td>
</tr>
</tbody>
</table>
5 Conclusion

Results show a correlation between the region of constriction of the acoustic tube and the place of articulation of the plosive. Nevertheless, we observe a strong variability with the vocalic context, which is not surprising given the simplicity of the defined cues. The acoustic tube has a complex dynamic behaviour, which cannot be accounted for by introducing such simple articulatory cues. The definition of context-dependent cues could achieve more accurate results.

Bibliography


XIIth INTERNATIONAL CONGRESS OF PHONETIC SCIENCES

Second Announcement

August 19-24, 1991
Aix-en-Provence, France

Institut de Phonétique
URA 261, CNRS : Parole et Langage
Université de Provence