Acoustic-Articulatory Inversion Based on a Neural Controller of a Vocal Tract Model: Further Results*

Alain Soquet  Marco Saerens  Paul Jospa
Institut de Phonétique and IRIDIA
Université Libre de Bruxelles CP 110
50, av. Franklin Roosevelt
B-1050 Bruxelles - Belgium

Abstract
We recently used the Distinctive Regions and Modes theory ([13]), coupled with a neural controller ([16]), to produce an acoustic-articulatory inversion of a vocal tract model. This paper presents further results for a 30-sections vocal tract. The inversion is now generalized to the whole vowel space.

1 Introduction

Mrayat, Carré & Guérin ([13]; 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 ([12]). It supplies relationships between the variations of the first three formants and the cross sectional areas of the eight vocal tract regions of the model. Previous work has shown that such a-priori knowledge can be used to control and invert non-linear physical processes with a neural network ([16]). 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 many-to-one 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]).

Preliminary results for vowels were reported in a previous paper [17]. It was shown that the network is able to learn to invert the process, for the eleven French oral vowels. The same experiment was performed with a constraint on the average volume of the vocal tract. This constraint allows the system to provide more realistic vocal tract shapes, and the convergence rate of the network is clearly improved. These results are now extended to a 30-sections vocal tract by introducing a continuity constraint, and the inversion is generalized to the vowel space.

Bailly et al. [2] are currently studying a similar, but more ambitious problem: They use Jordan's approach to control Maeda's articulatory model ([19]).

2 The Distinctive Regions and Modes theory

Vocal tract shapes are generated in the framework of the so-called Distinctive Regions and Modes theory ([13]; [3]). The model involves an acoustical tube closed at one end (glottis), and open at the other (lips) (Figure 2). This model is based on the study of acoustical properties of the vocal

*We would like to thank René Carré for helpful discussions, and Philip Miller for his help in the draft of the English text.

1The 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.
tract shapes, compared to those of a neutral uniform tube. The study showed that it is useful to subdivide the vocal tract into regions of different length, the distinctive regions. For the three formants model, eight distinctive regions are defined. Varying the mean cross sectional area of each of these regions induces specific and quasi monotonic formant variations. This behaviour can be qualitatively summarized by the signs of the formant variations associated with each region. The four front regions are anti-symmetric with the four back ones and vice versa. The eight regions will be denoted as $\sim A, \sim B, \sim C, \sim D$, and $D, C, B, A$. Furthermore different acoustic-articulatory modes can be distinguished, according to the constriction degree of a certain region. Vocalic sound productions can be associated to a large extent with the One Tract Mode that has the same qualitative behaviour as the neutral tract. When the degree of constriction of a certain region becomes too narrow, the behaviour of the vocal tract can no longer be approximated by the One Tract Mode.

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 1). Standard back-propagation cannot be used directly for the controller because the optimum control parameters ($u_0^*$, the optimum vocal tract cross sectional areas) are not known. An alternative is to compute the error at the output of the process. In order to perform the gradient descent relatively to this error, the Jacobian of the process, i.e. the partial derivatives of the output of the process ($y_j$) in terms of the control parameters ($u_i$), is necessary ([6]; [16]; [14]). For instance, Jordan ([5]) identifies the process with a neural network. Back-propagation is then performed through this net in order to compute the elements of the Jacobian ([4]). However, previous work has shown that, in some cases, the coefficients of the Jacobian can be approximated by their signs in the computation of the error at the output layer ($\delta_j$), without degradation in performances ([15]; [16]). If unit $j$ is an output unit, we find that $\delta_j$ is given by:

$$
\delta_j = f'(net_j) \sum q \text{ sign}(\frac{\partial y_k}{\partial y_j}) (y_k - y_k^d)
$$

(1)

where $f'$ is the derivative of a logistic activation function, $net_j$ the net total input of unit $j$, and $y_k^d$ is the desired output of the process (in this case, the target formant values). For hidden units, the standard rule is used. In our case, the signs are provided by the Distinctive Regions and Modes theory (the One Tract Mode).

![Figure 1: Architecture of the system.](image-url)

To perform the inversion (see Figure 1), the following three steps are iterated until the vocal tract model produces the target formants:

1. The neural network is given the 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 adapt the connection weights of the network with back-propagation algorithm (Equation (1)). If the network converges, the average number of iterations to reach the target formants must decrease towards one. In other words, the neural network learns to supply the correct cross sectional areas for the production of the target formant values.

4 Experiments

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. 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 2). 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. The learning parameters are fixed at \( \eta = 0.05 \) (the learning rate) and \( \alpha = 0.2 \) (the momentum term). We consider that the target formants are reached when \( |F_i^T - F_i^{\text{val}}| < 30 \text{ Hz, with } i = 1, 2, 3. \)

![Figure 2: Vocal tract divided in 30 parts and 8 different regions.](image)

Experiment 1: Minimisation of the cost function with a constraint on the average volume of the vocal tract and on the continuity:

\[
E = \sum_{v=1}^{3} \left( F_v^T - F_v^{\text{val}} \right)^2 + k_1 \left[ \sum_{i=1}^{30} L A_i^T - V_0 \right]^2 + k_2 \sum_{i=1}^{29} \left[ A_i^T - A_{i+1}^T \right]^2
\]

where the \( F_v^T \) are the formant values computed through the transfer function of the tube ([8]), the \( F_v^{\text{val}} \) are the target formants, \( L \) is the length of a part, the \( A_i^T \) are the corresponding areas supplied by the network, \( k_1 = 5 \times 10^{-5} \), \( k_2 = 2 \times 10^{-3} \), and the average volume \( V_0 = 55 \text{ cm}^2 \). The gradient descent is computed with equation (1). Ten runs have been performed, with different initial weights. The data set consists in the three first formant values of the 11 French oral vowels (we use values published by [10]).

As can be seen in Figure 3, initially, the net needs several iterations to reach the target (an iteration corresponds to one back-propagation). The number of iterations converges gradually towards one, which means that the network learns to supply the correct cross sectional areas to produce the vowels.

Experiment 2: Generalization to the vowel space:

The first two target formants are now uniformly generated in the vowel space (see Figure 4) and the third formant in the range [2400Hz, 2900Hz]. The values of the parameters are the same as in the first experiment. Results averaged over ten runs are presented in Figure 5.

Shapes corresponding to the eleven French oral vowels ([10]) are presented in Figure 6.
5 Conclusion

This paper presents results for an acoustic-articulatory inversion of an 30-sections vocal tract. By introducing two constraints, one on the average volume of the vocal tract, and another to ensure continuity between the different sections, we were able to generate vocal tract configurations that produce target formants. After the training, the network approximates the nonlinear mapping from the acoustic parameters (the three first formants) to the articulatory space (the cross sectional areas). Since it is a many-to-one problem, the network provides one possible solution to this problem. The constraints are introduced in order to reduce the number of possible solutions. Hence, we observed that the different mapping that were obtained with different initial weights are quite similar. Future work will be directed towards the study of the transitions between vowels, and the behaviour in the neighbourhood of consonant sounds ([19]).
Figure 6: Vocal tract shapes corresponding to the eleven French oral vowels.
References


reprinted from

ARTIFICIAL NEURAL NETWORKS

Proceedings of the 1991 International Conference on Artificial Neural Networks (ICANN-91)
Espoo, Finland, 24-28 June, 1991

Edited by

Teuvo KOHONEN
Laboratory of Computer and Information Science
Helsinki University of Technology
Espoo, Finland

Kai MÄKISARA
Instrument Laboratory
Technical Research Centre of Finland
Espoo, Finland

Olli SIMULA and Jari KANGAS
Laboratory of Computer and Information Science
Helsinki University of Technology
Espoo, Finland

1991

NORTH-HOLLAND
AMSTERDAM • NEW YORK • OXFORD • TOKYO