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 controller ([15]), to produce an acoustic-articulatory inversion of a vocal tract model ([17]). This paper presents results on the articulatory inversion of a vocal tract model plosives on the basis of this inversion scheme. We 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],[11],[7],[11]).

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],[11],[7],[11]).

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.

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]).

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.

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:

\[ E = \sum_{i=1}^{n} \left( \sum_{j=1}^{3} (F_i^j - F_{i0}^j)^2 \right) + k_1 \left( \sum_{i=1}^{n} L A_i^j - V_0 \right)^2 + k_2 \sum_{i=1}^{n} (A_i^j - A_{i0}^j)^2 \]

where the \( F_i^j \) are the formant values computed through the transfer function of the tube ([8]), the \( F_{i0}^j \) are the target formants, \( L \) is the length of a part, the \( A_i^j \) are the corresponding areas supplied by the network, \( k_1 = 5 \times 10^{-5}, k_2 = 2 \times 10^{-5} \), 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, Mrayati, Carré & Guérin ([14]) 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 Macspeech 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 × 6 × 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 Macspeech Lab software, at two different locations, for both adjacent vowels. The middle of the stable portion of the vowel (\( t_a \)) and the end of the vocalic transition (\( t_t \)) were used 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 \( t_a \), and a dynamic cue, which is:

\[
I(t_a) = \frac{A(t_a) - A(t_0)}{A(t_0) - A_{min}}
\]

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