INTRODUCTION
Acoustic-to-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 (Mermelstein, 1967; Atal, Chang, Mathews & Tukey, 1976; Larrar, Schroeter & Sondhi, 1988).

Previous work (Bailey, Bach, Laboliste & Olesen, 1990; Shirai & Kobayashi, 1991; Rahim & Goodyear, 1989; Rahim, Kleijn, Schroeter & Goodyear, 1991; Soquet, Saerens & Jospa, 1990; 1991) has shown that neural networks can be used in the framework of acoustic-to-articulatory inversion problems. In this paper, we present results obtained with different approximations of the sensitivity functions of the vocal tract, which are necessary for the computation of the back-propagated error. We compare the resulting vocal tract shapes to those published by (Majid, Boe & Perrier, 1986), and by (Fant, 1960). A new method for the approximation of the gradient descent, based on a variational formulation, is also introduced.

LEARNING ALGORITHM
A gradient descent algorithm, initially designed for adaptive control with a multilayer perceptron (Saerens & Soquet, 1991), can be applied to the control of a vocal tract model (Soquet, Saerens & Jospa, 1990; 1991). Let us briefly introduce the algorithm (Figure 1). For a p-order process $y_0(k) = a_0 y_0(k-1), ..., y_p(k-p), u(k-1), ..., u(k-p)$, where the $y_d(k)$ are the outputs of the plant at time $k$, and the $u(k)$ are the control parameters at time $k$, the gradient can be computed by introducing the standard Lagrange function $L$:

$$L = \frac{1}{2} \sum \frac{1}{\alpha} (y_d(k) - y_d^d(k))^2$$

$$+ \sum \Gamma \sum \frac{1}{\alpha} (y_d(k) - g_a(y_d(k-1), ..., y_p(k-p), u(k-1), ..., u(k-p)))$$

$$+ \sum \left( \sum \frac{1}{\alpha} (u(k) - U^d_a(k)) \right)$$

$$+ \sum \left( \sum \frac{1}{\alpha} (u_d(k) - U^d_a(k)) \right)$$

where $U^d_a(k)$ is the activation of unit $a$ at layer $l$ and time step $(k)$, and $w^d_a(k)$ is the weight of the connection between unit $a$ at layer $l$ and unit $b$ at layer $l-1$, also at time step $(k)$. The first term of equation (1a) corresponds to the cost function (the squared difference between plant output vector and target output vector), the second one (1b) to the input-output representation of the process of order $p$, the third one (1c) imposes that the units of the output layer $U^d_a(k)$ (layer $q$) provide the control parameters $u_d(k)$ at each time step $k$, and the last one (1d) corresponds to the transfer functions of the units. $\Gamma, A(q), x(q)$ are Lagrange multipliers.

Figure 1: Learning architecture.

The inputs of the net are the desired outputs $y_d^d(k+1)$ of the plant, the $p$ last output vectors of the plant, as well as the $p$ last control parameter vectors, in order to be able to reconstruct the state of the plant.

The extremum conditions are:

$$\frac{\partial L}{\partial u(k)} = 0, \frac{\partial L}{\partial \Gamma} = 0, \frac{\partial L}{\partial \Gamma_0} = 0,$$

and allow us to compute the Lagrange multipliers $\Gamma, A(q), x(q)$ (Le Cun, 1989). The weights are obtained by performing a gradient descent on $L$ (for more details, see Saerens, 1991; Saerens & Soquet, 1991).

In fact, the computation of the gradient involves the Jacobian of the process $\frac{\partial y_d(k)}{\partial u(k)}$. In our case, the elements of the Jacobian can easily be derived from the sensitivity functions.

ACOUSTIC-TO-ARTICULATORY INVERSION
Vocal tract shapes are generated in the framework of the so-called Distinctive Regions and Modes theory (Mirayi, Card & Guerlin, 1988). The model involves an acoustic tube closed at one end (glottis), and open at the other (lips) (Figure 2). It is based on the study of the acoustical properties of vocal tract shapes, compared to those of a neutral uniform tube. For the three formants model, eight regions of different lengths (the distinctive regions) can be defined. Varying the cross sectional area of each of these regions induces specific and quasi monotonic formant variations.

A multilayer 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 3). The algorithm described in previous section is used for the training of the net.

Figure 2: Acoustic tube divided in eight regions of different lengths.

Training is realized on the whole vowel space. By introducing a constraint on the volume of the acoustic tube, the network converges without problem (Soquet, Saerens & Jospa, 1990; 1991; Saerens, 1991). Moreover, the first section (\(-A\)) is limited to the range [1.0 cm$^2$, 2.5 cm$^2$]; the
other sections to $[0.5 \text{ cm}^2, 15.0 \text{ cm}^2]$. The optimized cost function, corresponding to (1a), is:

$$E = \frac{1}{2} \sum_{i=1}^{8} (y_{i} - \hat{y}_{i})^2 + \mu \sum_{i=1}^{3} (\sum_{j=1}^{l_{i}} A_{ij}^2 - V_{0})^2$$

where the $F_{ij}$ are the formant values associated with the acoustic tube; the $F_{i}^{wd}$ are the target formant values corresponding to vowel $v$; the $L_{j}$ are the lengths of the different regions; the $A_{ij}$ are the cross sectional areas provided by the network; $V_{0}$ is the total volume of the tube; and $\mu$ is a constant value set to $5 \times 10^{-5}$. The summation is carried out on the vowels $v$ generated in the vowel space.

The learning process is as follows (Figure 3):

1. Target formants $F_{i}^{wd}$ are generated in the vowel space. These values are given as input to the network.
2. The outputs of the network supply the vocal tract cross sectional areas $A_{ij}$. The corresponding formant values $F_{ij}$ are computed (we use the algorithm published by Liljencreutz & Fant, 1975).
3. The difference between these formant values and the target formant values are used to adapt the weights of the network. Go to step 1 until convergence.

Figure 3: The network is trained in order to provide the cross sectional areas that produce target formants.

After training, the network approximates the nonlinear mapping from the acoustic parameters (the three first formants) to the articulatory ones (the cross sectional areas). Since it is a one-to-many problem, the network provides one possible solution to this problem. The penalty on the volume is introduced in order to reduce the number of possible solutions. Hence, we observed that the different mappings that were obtained with different initial weights are quite similar (Sorens, 1991).

APPROXIMATION OF THE JACOBIAN

We investigated two different approximations of the Jacobian values, used for the computation of the gradient: a fixed approximation and a perturbation approach. However, as the computation of the perturbation is time-consuming, we propose to compute precisely the sensitivity functions by using a variational formulation (Jospa, 1972, 1977). This algorithm is more economical in computation time.

1. Fixed approximation

In this first case, the elements of the Jacobian of the model are kept constant for every vocal tract configuration. The values are chosen so as to reflect the qualitative behaviour of the vocal tract model as described in (Mrayati, Carre & Guérin, 1988).

2. Perturbation approach

In this second case, the values of the Jacobian are updated according to the configuration of the vocal tract model. For each configuration, every control parameter is slightly modified and the corresponding variations of the formant values are used to approximate the Jacobian elements. This method is more appropriate, but is time-consuming.

3. Explicit computation of the sensitivity functions

In this last case, the values of the Jacobian are also updated according to the configuration of the vocal tract model. The sensitivity functions are computed thanks to a variational formulation of the wave propagation problem in the vocal tract (Jospa, 1972, 1977; 1982). This interesting approach is less time-consuming, and gives an accurate approximation of the sensitivity functions, which can then be used to compute the elements of the Jacobian.

Within the framework of the adopted acoustical model (Sondhi, 1974), the amplitude distribution $w_{n}(x)$ and the frequency $f_{n}$ of the resonance mode $n$ of the vocal tract are linked to the area function $A(x)$ by the following equation:

$$\frac{\partial}{\partial x} (A \frac{\partial w_{n}}{\partial x}) + \frac{4 \pi^{2} f_{n}^{2}}{c^{2}} (\int_{0}^{L} A(x) \ dx) \ w_{n} = 0$$

where $c$ denotes sound velocity, and $f_{n}$ denotes the effective natural resonance frequency of the tract walls ($f_{n} = 2000/n$).

We impose:

$$\partial w_{n}(0) = 0 \ , \ \text{at glottis} \ (x=0),$$

$$\partial w_{n}(L') = 0 \ , \ \text{at lips, with} \ L' = L + 0.66 \ \sqrt{\overline{A}(L)},$$

where $L$ is the vocal tract length. The area function is defined by a set of articulatory parameters $\{a\} : A(x) = A_{e}(x; \{a\})$ (an articulatory model). The equation (3), with boundary conditions (4a) and (4b), can be considered as the Euler-Lagrange conditions that $A$, $w_{n}$, and $f_{n}$ must satisfy to make the functional (5) stationary with respect to the variations of $w_{n}$ ($\delta \left[ w_{n} \right] = 0$):

$$L = \int_{0}^{L} \left[ (\partial_{x} w_{n})^{2} - \frac{4 \pi^{2} f_{n}^{2}}{c^{2}} \ w_{n}^{2} \right] \ dx$$

Application of the Rayleigh-Ritz method to this variational problem leads to conditions (Jospa, 1972, 1977, 1982) that allow the computation of the formants of the vocal tract, and gives explicit estimations of the sensitivity functions. This is an efficient way to find the formant values, as well as the sensitivity functions, corresponding to a given vocal tract. This algorithm has already been used in (Jospa, 1991) for acoustic-to-articulatory inversion with a standard optimization method. We plan to use it instead of the algorithm of (Liljencreutz & Fant, 1975), used so far. Unfortunately, we do not have simulation results yet for this method.

RESULTS AND CONCLUSION

The resulting vocal tract shapes are provided by a network after training on the whole vowel space. We use a different network for each vocal tract length and for each Jacobian approximation. Figure 4 shows vocal tract shapes obtained for french vowels (the formant values are from Majid, Boe & Perrier, 1986). The first left shape (original), corresponds to the configuration published by Majid, Boe & Perrier, 1986). The second shape (signs) was obtained by using fixed values for the Jacobian. These values come from the study of (Mrayati, Carre & Guérin, 1988). The length of the acoustic tube is 19 cm. The two next vocal tract shapes were obtained by evaluating the Jacobian at each iteration with small perturbations. The third shape corresponds to a tube of 19 cm and the fourth to a tube of 17 cm.
Figure 5 shows the vocal tract shapes obtained for the russian vowels of (Faut, 1960). The left shape is the one published by (Faut, 1960). Only the results obtained by using perturbations are represented, for three different lengths: 19 cm, 18 cm and 17 cm.

Comparisons of the shapes obtained by the two different methods show that the fixed approximation introduces some unrealistic behaviour for closed vowels like [u] and [o]. This result can be explained by the fact that, for this kind of vowel, the behaviour of the acoustic tube is closer to a tube closed at lips (Carré & Marayat, 1991). The use of an adaptive approximation seems to solve a great deal of the problems encountered with the closed vowels. The application of the Rayleigh–Ritz method, briefly described below, should give comparable results, with less computation time.

REFERENCES


Figure 4: From left to right: the reference published by Majdi, Boë & Perrier (first figure); vocal tract shape provided by the net — training with constant values for the Jacobian (second figure); and by making small perturbations (for two different acoustic tube lengths; third and fourth figures).

Figure 5: From left to right: the reference published by Font (first figure); vocal tract shape provided by the net — training by making small perturbations (for three different acoustic tube lengths; second, third and fourth figures).
SIGNAL PROCESSING VI
THEORIES AND APPLICATIONS