A Gesture Interface Controlled by the Oral Cavity

Nicola Orio
CSC, Dipartimento di Elettronica e Informatica, Università di Padova
orio@dei.unipd.it http://www.dei.unipd.it/english/csc

Abstract
In this paper it is proposed a new gesture interface which reveals and maps postures of the oral cavity: the user has only to change the inner shape of the mouth. The interface puts a white noise inside the oral cavity and picks the filtered signal up with a system of two acoustical waveguides. The oral cavity filter parameters are estimated with an analysis of the recorded signal. Then the Principal Component Algorithm is used to extract independent parameters which are used for control. The interface needs a short training phase in which a new user assumes all the postures he/she wants to use for the control. Even if the interface was designed for musical application, it can be used also in other fields; for example as a mouse for people with disabilities, or as an aid tool in phonetics. A prototype of the interface was built. It gives a three-dimensional output value, on a MIDI device, every 0.1 seconds.

1 Introduction
It is a widespread opinion that playing an electronic music instrument gives artistic results not comparable with those obtained with a common acoustic instrument: there is a gap between the expressivity reached with the new musical instrument and the one reached with the classical ones. Since the act of playing consists essentially in producing a sound and continuously controlling its characteristics in timing and in timbre, a performance depends mainly on the interaction degree between the player and the instrument. So, also considering that sound synthesis can potentially give a great variety of timbre nuances, this gap is not due to intrinsic limitations of the instruments but to the control musicians have on them: there is still a weak interaction and a low amount of information the musicians can exchange with the sound synthesis algorithm.
In this paper it is proposed a new gesture interface which can be used while playing a MIDI instrument, so to increase the amount of data the musician can give, in real time, to the synthesis algorithms.

2 The Oral Gesture Interface
The proposed oral gesture interface maps oral cavity postures in a space: points of this space are the output of the interface. This mapping operation can be divided in three main steps: man-machine interaction, parameters extraction and data processing. The first step is represented by the physical interface, which reveals the postures of the mouth: its output is a sampled audio signal $s(k)$ which is correlated to the shape of the oral cavity. In the second step the signal is analyzed: using well known techniques of digital signal processing a set of parameters is extracted. The output of the parameters extraction step is a vector $p(k)$ in the discrete-time domain. The third step consists in transforming $p(k)$, which belongs to the parameters space and has correlated elements, to a vector $e(k)$ belonging to the control space with independent elements; so $e(k)$ is a multidimensional signal in the discrete-time domain, in particular it can be a vector of MIDI signals.

In figure 1 it is shown a complete scheme of the interface.

2.1 Man-Machine Interaction
For revealing postures the oral cavity can be considered like a resonant tube with time-varying inner shape, i.e. a linear filter in the audio frequency range: estimating filter parameters means to have the required information on the shape of user's mouth. This reduces to a deconvolution if the input signal is known.
Since one of the aims is that the user does not have to produce sounds, the interface provides the signal which will excite the system. To obtain this, it is proposed the use of two physical waveguides; the first, having a small loudspeaker on its outside end, puts the signal inside the oral cavity and the second, which has a microphone, picks up the signal: the picked signal is the result of filtering made by the oral cavity, the comparison by deconvolution between the input and the output signals gives the needed information on the mouth shape. The two waveguides have to be thin enough to be very little troublesome for the user, practically they have to reach the mouth like the mouthpiece of a reed instrument, e.g. a clarinet.
Some studies were made on the kind of signal to be put inside the oral cavity. They were tested a white noise, square, triangle and saw tooth waves and with a pulse train. Best results in terms of control were obtained with the white noise. Furthermore this signal is completely masked by sounds of the performance and by background noise so the musician does not have the annoyance of hearing a sound inside the mouth.
2.2 Parameters Extraction

To extract the parameters of the sampled signal \( s(k) \) it is necessary to perform a deconvolution. There are many algorithms developed for audio signals, in particular for speech recognition. Since the interest is upon a real-time use of the controller, we have to consider algorithms which give good results with a low sampling rate and with a few input samples.

One of these is the Linear Prediction Method [1][2]. As it is known LPC extracts, with a recursive algorithm, the polynomial coefficients of an all-pole filter minimizing the error between the sample at time \( k \) and its prediction from previous samples. Practically, LPC performs a deconvolution of the signal, thus giving a representation of the filter in terms of polynomial coefficients.

An iterative algorithm was adapted from LPC autocorrelation method, in order to extract the reflection coefficients of the resonance tube which approximates the oral cavity. There is a certain freedom in the choice of the filter order and hence in the output vector dimension. This value is linked to the sampling rate and also with the geometry of the oral cavity. After some tests it was decided to choose to extract 12 parameters, because there was no noticeable difference from the control ability with order 12 or 14. On the other hand a filter order smaller than 10 reduced the control dimensionality.

To see if the results were independent by the method used for the data extraction, another set of parameters was obtained from the sampled signal: the Linear Frequency Cepstral Coefficients [3]. All the tests were made on both the sets of coefficients and the results were quite similar. However, the LPC was chosen for building the prototype of the interface.

2.3 Data Processing

It is evident that extracting the filter parameters does not mean to obtain an effective control. To this end, it is necessary that the outputs are naturally independent. Therefore it is necessary to extract a vector with independent elements which carries the same information of the parameters vector. The problem may be expressed in this way: *given a vector \( p(k) \) with \( M \) correlated elements, find \( N \) linear combinations of the \( M \) variables (with \( N < M \)) forming a vector \( c(k) \) with \( N \) independent elements which carries the same information of \( p(k) \).*

Determining the value of \( N \) means, in this case, determining the dimension of the control space. A solution is given by the Principal Components Algorithm [4][5]. The PCA starts with a set of samples of the \( M \)-dimensional vector; from these samples a projection matrix is built which transforms the parameters space into the control space. This means that, before use, the oral gesture interface needs a training phase during which each new user has to give all the information needed to calculate the principal components. Practically, during the training phase the user has to assume all the postures he/she can use for the control; the interface samples the parameters evaluated from these postures, and with PCA it evaluates the number of independent linear combinations, that is the dimension of the control space, and the projection matrix. The need of a training phase has some drawbacks and some advantages. The main drawback is the increased complexity of the software support for the gesture.
interface; it requires more memory and also numeric algorithms for evaluating the projection matrix are required. On the opposite, the main advantage is that the controller takes a new configuration for each new user, hence everyone has a complete control of the output combinations.

3 Analysis of Results

First of all, it was seen that the dimension of the control space can be considered to be 3: it is a quite satisfactory result, considering that the dimension of standard MIDI controllers is lower. Others components have to be ignored because they are significantly affected by noise.

An analysis was made on output firmness of \( e(k) \) parameters of the control space when the user assumes the same posture for long time (10 secs). There were some small oscillations of the principal components around an average value which is about 8-10%; this can be explained both with involuntary movements of the user and with the presence of numeric approximation.

A next step in data analysis was to test if, for each point inside the control space, there exists at least a posture which maps on it. There were no area inside the control space which is not reachable by the interface outputs, apart near boundary values (see figure 2). This is easily explained considering the particular choice of the output mapping and that PCA maps the elements of the output vector inside an ellipsoid.

Another analysis was made upon the trend of the values relating to some movements. Practically, it is desirable that, for particular movements of the mouth that are natural for a user, outputs have a trend regular enough to be recognized, and it was noted that there are some movements which map in monothonic trends of the outputs.

Another kind of analysis was about the characterization of the link between the principal components and the spectrum of the signal taken by the oral cavity. The principal component are calculated using the Linear Frequency Cepstral Coefficients, because in this case the effect of each component on the signal can be summed, while there is no linear relation using the reflection coefficients. In figure 3 are plotted the first three base vectors. It can be seen that the first base vector is related to a low-pass characteristic of the signal and it has a formant at 1 kHz. The second one has two antiformants, one at 1.5 kHz and the other at 3.2 kHz. Considering the mixed effect in Fig. 11, it is interesting to note how the first component controls the energy at about 1 kHz while the second one controls the energy at low frequencies. The third base vector clearly controls two formants, at 800 Hz and 2.2 kHz, and one antiformant at 3 kHz.

4 Using the Interface

Training times taken by a new user were measured experimentally. To better estimate the difficulties in use the outputs were redirected to a graphic interface, in which three sliders were associated to the outputs of the controller. Learning tests were made at the beginning with only one output (the first) and, after the user succeeds in controlling it, augmenting the number of output. Tests were made with three subjects, only one of them was a musician, a flute player.

The control of one dimension is quite immediate. After a few minutes users learn how to move the slider in both directions and, after a little while, they learn to obtain absolute positions. Also controlling two dimensions is easy, especially after some practice with the mono-
dimensional control. In less than an hour it is possible to learn how to control the relative position of the two sliders. The control of absolute positions is a little harder, and it generally takes another hour. Controlling all the three dimensions, instead, is quite difficult: it takes some hours.

5 Methods for an Easier Use of the Interface

To overcome the difficulties in use of the controller, two methods are proposed.

The first one is based on the idea that, for a user, it would be simpler to control the outputs if they are correlated to some physical characteristic: for instance, the formants frequency. During the training phase, for each sampled posture, the first two formant are evaluated. The cross-covariance matrix between the formants frequency and the principal components is then built. It is possible to prove that the linear combination of the reflection coefficients with the first row of the cross-covariance matrix gives an output which has maximum correlation with the first formant; similarly the second row gives maximum correlation with the second formant, while the third component is evaluated to be orthogonal to the first two. After this transformation the user can refer to formant frequency while learning to use the interface, and the control will be more natural.

The second method consists in finding a rotation of the control space, so that some elementary movements will affect only one output component. After finding some movements which map in approximately linear trends of the outputs, a rotation matrix has to be built so that these particular movements will lie on the axis of the control space. This can be done using, again, a statistical method: the multivariate regression. This method is the multidimensional extension of the classical linear regression method; given two sets of sampled variables which describe the same object, it allows to linearly transform one set to the other. In this case the first set is the sampled trend of the principal components during the movements, and the second set is the wanted trend along the axes.

6 Applications

Until now we always refer to musical application of the oral gesture interface, for instance to control a synthesis algorithm in real time performances. But, after seeing the good performances of the controller, there may be also non-musical applications. First of all, since the controller does not require the use of hands, it can be used as a man-computer interface for people with disabilities subjects with limb paralysis. An immediate application may be the use like a mouse simulator: the first two principal components can be used to change the cursor position on the screen and another interface (i.e. a switch controlled by the lips) can be used to make the selection. After this, considering that the interface makes a map of oral cavity postures in a three dimension space, the controller can be an aid in phonetic studies. Probably it can be also used by deaf subjects for learning languages; it is known that learning a spoken language needs the continuos aid of a teacher, who help the subjects who can not ear themselves. Using the controller, the subject can learn to reproduce some postures associated to vowels or steady sounds only trying to obtain a particular output configuration.

7 Conclusions

A new gesture interface was built in which control is obtained with postures of the oral cavity. The prototype of this controller [6], implemented on a NeXT platform, gives a three dimensional output every 0.1 seconds. The controller can be used in parallel with a common MIDI instrument, and can integrate the information flown from a musician to a sound synthesis algorithm, so that a deeper level of interaction and an expressive performance can be achieved. Anyhow the controller can be used in all the fields where it is useful to have a mapping of the oral cavity postures in a three dimensional space with independent components.

References