CALIBRATION OF PHYSICAL MODELS USING ARTIFICIAL NEURAL NETWORKS WITH APPLICATION TO PLUCKED STRING INSTRUMENTS

Ali Taylan Cemgil(1), Cumhur Erkut(2)

(1) Department of Computer Engineering, Boğaziçi University TR-80815 Istanbul Turkey
cemgil@boun.edu.tr
(2) Department of Electronic and Communication Engineering, Yildiz Technical University, TR-80670 Istanbul, Turkey
erkut@yildiz.edu.tr

ABSTRACT

This study introduces a method for estimating the model parameters of plucked string instruments using artificial neural networks (ANN's). The system consists of a preprocessor and an ANN model. The preprocessor computes the time-varying harmonics of recorded sounds and this data is fed into the ANN model to implement the nonlinear mapping to the parameter space. The main advantage of this method is that it performs the mapping according to a perceptual measure derived from listening tests.

1. Introduction

The use of physical models to synthesize musical sounds in real-time has brought new concepts to sound synthesis. As opposed to sampling synthesis, which provides only a few samples from a broad timbral space, physical modeling focuses on the sound production mechanisms, thus aiming to describe the musical sound by a small set of intuitive and flexible parameters. Researchers consider precise physical relations (often limited by the hardware available) as well as the properties of human auditory system to exclude perceptually indiscriminable effects to obtain efficient schemes.

For synthesizing plucked string tones, a very efficient algorithm has been developed by Karplus and Strong [1]. Jaffe and Smith [2] formulated the algorithm as a simulation of a physical system with well defined input-output relations and extended the basic scheme by introducing various filters in the feed-back path to simulate some of the observed physical properties of string instruments. Sullivan further extended the basic algorithm to synthesize electric guitar sounds [3]. Smith later generalized this ideas to formulate digital waveguides, and using digital waveguides he obtained a general physical modeling scheme [4]. Recently, Valimaki et. al. [5] presented an overview on model-based synthesis of plucked string instruments.

In general, the parameters of these models are calibrated from the recordings of traditional musical instruments using some parametric and nonparametric DSP procedures [5], [6]. These procedures, however, do not guarantee a perfect match between the original sound and the model output in a perceptual sense. Besides that, the quality assessments about the final model output can not be directly fed back to the parameter estimation procedure.

The present work introduces a method to calibrate the physical models using perceptual criteria. The outline of the method is as follows: i) Decomposition of the model into basic blocks controlled with the minimum number of parameters, which need to be calibrated according to a perceptual measure. ii) Obtaining the useful range of the submodel parameters with listening tests. iii) Determination of the submodel output sensitivity to the change of parameters with listening tests. iv) Interpretation of the results from the listening tests to obtain a perceptually compatible metric. v) Calibration of the submodel parameters with an Artificial Neural Network (ANN) estimator using the metric of step iv.

As a case study, we constructed a simple submodel for simulating plucked string tones. The model consists of a delay line, a simple loop filter and a tuning filter. The loop filter $H_l(z) = \frac{1-a}{1-az^{-1}}$. is proposed in [2] and used with minor modifications in [5]. The all-pass tuning filter $H_t(Z) = \frac{C+z^{-1}}{1+Cz^{-1}}$ controls the overall phase delay [2]. In this simplified model (Fig. 1) there is just one parameter to control the timbral variation; the loop filter coefficient a.

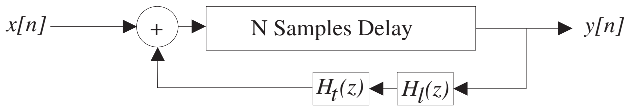


Figure 1: Simplified Plucked String Model

2. Listening Tests

The tests were carried out using a program with a graphical interface ,which synthesize the test sounds and process the listener responses. We used a fixed-seeded white-noise sequence $x[n]$ to excite the model for equalizing the timbral dependence of the model output to the excitation sequence.

2.1 Determination of Subjective Distances

We conducted a listening experiment to obtain the useful range for a as a function of frequency. By a useful range we mean the set of parameters that generate sounds which indeed "sound" like a plucked string instrument rather than just a click or a buzz. Three musically trained students of Yildiz Technical University (YTU) are asked to classify sounds generated at different frequencies with randomly generated a's around the prior boundary, that we subjectively set. An exponential curve fit to the test results gives the upper $(a_h(f))$ and lower $(a_l(f))$ boundaries of useful region.

The perceptual difference between sounds can be assessed by conducting a series of listening tests in which the subjects are asked to rate the differences between sound pairs generated by the model. We take seven equally spaced samples from each $[a_l(f), a_h(f)]$, resulting 49 pairs among each frequency. In this work, we selected four frequencies f_k beginning with $f_1 = 110Hz$, where $f_k = 2^{k-1}f_1, \quad k = 2, 3, 4$. Five musically trained YTU students were asked to rate the distances between this pairs for each f_k. The rating is done on a 4 level scale with the following interpretation: 0) The sounds are the same, 1) Slightly different but percievable only by paying attention, 2) The difference is easly percievable and 3) Obviously different sounds.

The results of such an experiment can only give *information about* the underlying true "perceptual metric"; indeed only an ordering of the representative sounds can be obtained in the space of all possible sounds that the model can generate. However, if we assume that the space is smooth, then this ordering information could be used to estimate a perceptually compatible metric between the sounds by multidimensional scaling methods.

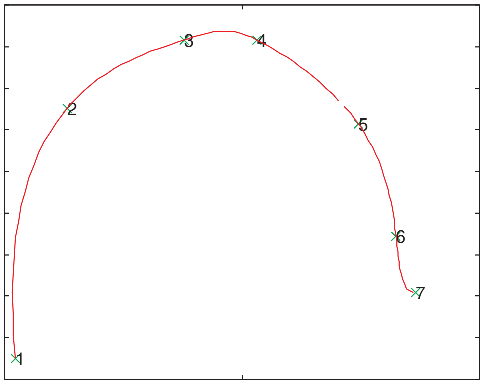
2.2 Multidimensional Scaling

The term *multidimensional scaling* refers to the collection of techniques for finding a configuration $\mathbf{X}$ of points (preferably in 2 or 3 dimensions) given the mutual distances or just a partial ordering information about these distances. This is basically the inverse process of computing the pairwise distances $\mathbf{D} = \{d^2_{rs}\}$ between a set of points [7], [8].

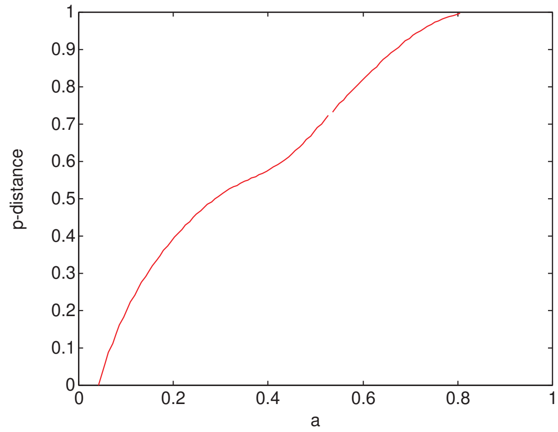
Figure 2 shows the scaling of the ordering information obtained from a listening test. The points arrange themselves in such a way that it is possible to fit a smooth parametric curve with the loop filter coefficient as the free parameter.

2.3 Perceptual Distance

Once a configuration $\mathbf{X}$ is found on the map, the arc-length on a smooth curve that connects the particular points x_i can be used to estimate the "perceptual distance" between the sounds. In this work we employ



a)



b)

Figure 2: An Example of Multidimensional Scaling of ratings(a) and derived perceptual distance (b)

the well known cubic spline interpolation method.

The arclength $s_f(a_1, a_2)$ of the spline interpolant $\mathbf{q}$ joining two points in N dimensions for a fixed frequency f is given by

$$s_f(a_1, a_2) = \frac{1}{S_f} \int_{a_1}^{a_2} \sqrt{\sum_{i=1}^{N} \left(\frac{dq_i(\tau)}{d\tau}\right)^2} d\tau$$

where q_i is the spline fit to the i'th dimension of $\mathbf{X}$, $S_f = \int_{a_l(f)}^{a_h(f)} \sqrt{\sum_{i=1}^{N} \left(\frac{dq_i(\tau)}{d\tau}\right)^2} d\tau$ is a normalization constant equal to the total length of the useful range. Note that the derivative of this curve with respect to a gives an estimate of the sensitivity of the ear. The computed distance measure as a result of listening tests is shown in Fig. 3. The constant frequency lines in the 3-D plot correspond to $s_f(a_l, \cdot)$ where the listening tests are carried out. At other frequencies, we find s_f by interpolation.

3. Estimation of Physical Model Parameters by an ANN

In general, we can view a physical model as a mapping $f : A^n \rightarrow O$ where A^n is the set of free parameters of the model and O is the set of sound signals that can be generated by the model. Thus, estimation of the physical model parameters can be formulated as an approximation problem to f^{-1}. Usually, f^{-1} is described as a DSP algorithm, which employes several heuristics and prior assumptions about the signal to be modelled. We can also take a slightly different approach and choose a general and flexible model for approximating f^{-1}. The approximation then has the form $\phi : g(O) \rightarrow A^n$ where $g(O)$ is the set of features (e.g. harmonic envelopes) extracted from the sound after some preprocessing $g(\cdot)$ and A^n are the model parameters (e.g. a).

The key issues are : i) Finding a suitable representation of the sound by preprocessing, ii) The choice of the model for ϕ.

3.1 Preprocessing

Computing the discrete STFT coefficients $Y(n, k) = \sum_{m=0}^{N-1} w(m)y(m - Ln)e^{-j2\pi km/N}$,

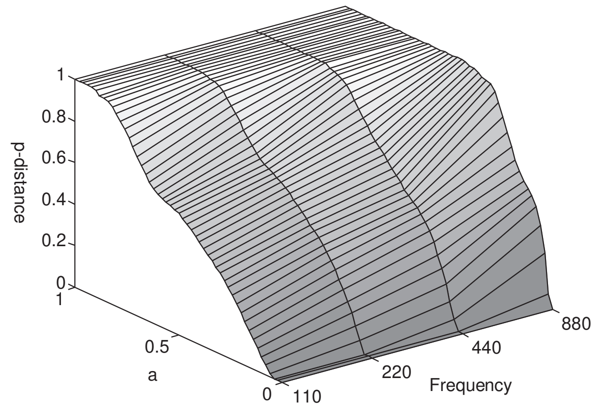


Figure 3: Perceptual Distance $s_f(a_l(f), a)$ derived from listening tests

we find a set of discrete 'harmonic envelopes' $y_h[n]$ from the STFT modulus in dB scale, where $h = 1 \dots H$ is the harmonic index and $n = 1 \dots R$ is the frame index. Here,$N = 2048$, w is a Hamming window, overlap $L = N/2$. Then we compute the first order difference sequence $\Delta_h[n] = y_h[n] - y_h[n-1]$.

3.2 MLP

In our study, we concentrate on a well known topology, the Multilayer Perceptron (MLP) which has the following functional form :

$$\phi = \mathbf{T}[\mathbf{\Psi}_t(\mathbf{W}\mathbf{\Delta})]$$

where $\mathbf{\Delta} = [\Delta_0, \Delta_1, \Delta_2, .., \Delta_p]$, a $p+1$ dimensional input vector with the constant bias term $\Delta_0 = 1$, where $p = HR$, $\mathbf{W}$ is a $(t-1) \times (p+1)$ input layer connection weight matrix, $\mathbf{T}$ a $n \times t$ hidden layer connection weight matrix, $\Psi_t(x) = [1, \psi_1, \psi_2, \dots, \psi_t]$ is a vector valued nonlinear sigmoid function where $\psi_i = \exp(x_i)/(1 + \exp(x_i))$. (See [9]). In our study, the output dimension $n = 1$ since we only estimate a.

3.3 Training

Once a topology for $\phi(\cdot, \mathbf{W}, \mathbf{T})$ is chosen, the connection weights are estimated by minimizing an error measure, which serves as an indicator of the distance between the desired output and the observed output obtained from the set of presented patterns to the network:

$$E = \sum_{i=1}^{M} d(y_i, \phi(\Delta_i, \mathbf{W}, \mathbf{T}))$$

where $d(\cdot, \cdot)$ is a distance measure and x_i are the ith training data sample from a set of M samples.

Conceptually, we desire that the perceptual difference between a sound and its model to be minimized. In general, this criteria is difficult to incorporate in a computational procedure since the perceptual difference can not be physically measured. Therefore, the error criteria is almost always taken as the Euclidean norm which assumes uniform variance in every region of the output space. This choice of error criteria may lead in certain instances to poor results since the perceptual difference is in general not a linear function in the set of parameters, i.e. the response of the human auditory system is not uniform in every region of the parameter space. Thus, the regions where the parameter has more impact on the resulting sound, its estimation "deserves" more precision. Because of this reason, we incorporate the perceptual distance in the error criteria. This is done by simply setting $d(\cdot) = s_f(\cdot)$.

Table 1: MSE ($\times 10^{-3}$) on test set : Mean and standard deviation of 5 independent runs for R frames and H harmonics.

			R	
		3	5	7
	5	1.24±0.22	0.96±0.17	0.80±0.10
H	6	1.33±0.13	1.14±0.24	0.84±0.06
	7	4.55±2.88	5.42±0.61	4.21±0.46
	8	7.82±1.80	4.45±2.19	3.14±0.78

Since the MLP model ϕ is non-linear in its weights, no close-form solution for the optimal parameters are known. Therefore a steepest descent algorithm is employed to estimate a "good enough" sub-optimal solution for $\mathbf{W}$ and $\mathbf{T}$. In order to avoid overfitting, during training the error on a cross validation set is monitored (e.g. on the test set), and training is stopped when this error tends to increase.

4. Results

The simulations are carried out using a training and a test set of 200 sounds generated by the model in Fig. 1 with random fundamental frequencies in the range 110 – 880 Hz. The inputs to the MLP are : i) Fundamental Frequency estimation (in Normalized Log-Scale), ii) Spectral Envelopes of H harmonics obtained by processing of R frames. The desired output is $s_f(a_l, a)$. The results are for a fully-connected network with 10 units in the hidden layer is summarized in Table 1.

The network output $s_f(a_l, a)$ can uniquely be mapped to the corresponding a. The increase in the number of frames results in better estimates because this provides more information about the decay. Increasing the number of harmonics makes the estimates of the preprocessor more unreliable since rapidly decaying higher harmonics reach the noise floor for $R > 6$. Fig. 4 demonstrates that the perceptual distance criteria helps to control the error in regions where the ear is more sensitive.

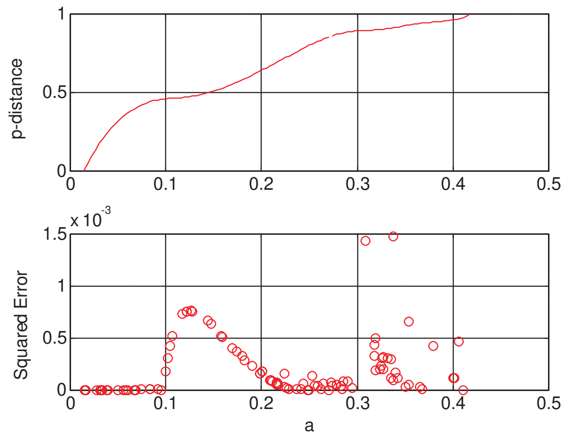


Figure 4: Perceptual Distance(i) and Squared Error(ii) as a function of a

In our own subjective tests we concluded that two sounds having a perceptial distance less than 0.05 are indistinguishable, thus the method provides useful results. In the simulations carried out with real data we observed that the method can approximate a real sound reasonably well especially the decay characteristics.

5. Conclusions and Future Research Plans

This relatively simple problem shows that ANN's can be used for calibration of physical model parameters. It is also fairly straightforward to include any perceptually compatible metric in this process. Presently, listening tests seem to be the expensive part of the method.

Another feature is the underlying generalization ability of ANN's. Using a conventional DSP algorithm, parameter estimation may be biased due to assumptions that may not hold for real data. A neural network is more robust to "failing assumptions" because it virtually extracts necessary features from the data along with the noise introduced by the preprocessor, thus is less biased.

ANN models can be used for estimation of other parameters of physical models efficiently (e.g. plucking point estimation). Different ANN topologies can also be used.

An important advantage of deriving the perceptual distance using the method described in this study is that the model can be analytically differentiated with respect to the free parameter, which can make optimizations analytically tractable. Additionally, if parameters are to be stored in a table, the use of such a perceptual criteria can help in a meaningful quantization of model parameters.

References

1. K. Karplus, A. Strong, "Digital Synthesis of Plucked String and Drum Timbres", Computer Music J. vol. 7, no. 2, pp. 43-55 (1983).
2. D. Jaffe, J. O. Smith, "Extensions of the Karplus-Strong Plucked String Algorithm", Computer Music J. vol. 7, no. 2, pp. 56-69 (1983).
3. C. R. Sullivan, "Extending the Karplus-Strong Algorithm to Synthesize Electric Guitar Timbres with Distortion and Feedback", Computer Music J. vol. 14, no. 3, pp. 26-37 (1990).
4. J. O. Smith, "Physical Modeling Using Digital Waveguides", Computer Music J. vol. 16, no. 4, pp. 74-87 (1992).
5. V. Valimaki, J. Huopaniemi, M. Karjaleinen, Z. Janosy, "Physical Modelling of Plucked String Instruments with Application to Real-time Sound Synthesis", J. Audio Eng. Soc., vol. 44 no. 5 pp. 331-353 May 1996
6. J. Laroche, J. L. Meillier, "Multi-channel Excitation/Filter Modeling of Percussive Sounds with the Applicaiton to the Piano", IEEE Trans. Speech Audio Process., vol. 2, pp 329-344 (1994 Apr.)
7. C. Chatfield, A. J. Collins, Introduction to Multivariate Analysis, Chapman and Hall, New York 1989
8. J. Grey, Multidimensional perceptual scaling of musical timbres, J. Acoust. Soc. Am. 61 (5), pp1270-1277, May 1977
9. B. D. Ripley, Neural Networks and Related Methods for Classification, J. R. Statist. Soc. B(1994) 56 No. 3, pp. 409-456