ACOUSTIC-TO-ARTICULATORY NEURAL MAPPING UNDER DIFFERENT STATISTICAL CHARACTERISTICS OF ARTICULATORY PATTERN VECTORS

H Altun* and K M Curtis≅
RISN Group, University of Nottingham
**Department of Electrical & Electronic Engineering, Nigde University, Nigde, Turkey*
≅*Department of Electrical & Electronic Engineering, University of Nottingham, Nottingham, NG7 2RD, UK*

ABSTRACT

This paper describes a mapping problem that tests and validates the findings from our analytical analysis of neural learning[1]. In this analysis different statistical characteristics of the target pattern vectors were investigated as to their effect on learning and generalisation. The problem reported on is a difficult function approximation problem, where the parameters of an articulatory speech synthesiser are estimated. The estimation of the articulatory parameters from the acoustic domain is difficult, due to the non-linear and ill-posed nature of the relationship between the acoustic and articulatory parameters. Despite proposals employing neural networks for this task in the past, neural network mapping has not been shown to be superior to other techniques that have been used to try and solve this inversion problem. Using the theoretical results of [1] we show that NN mapping can be used successfully to improve the solution to this problem.

1. INTRODUCTION

In speech synthesis, there is a consensus among researchers that the articulatory speech synthesiser has the potential to be the ultimate solution to the synthesis of natural sounding, intelligible speech. It promises greater naturalness and allows for a greater flexibility in adjusting to individual speakers [2, 3, 4]. Although much research has been carried out towards this end, the problem of estimating the control parameters for the articulatory synthesiser, from a given speech signal, remains still a challenge which has not been fulfilled completely to date [3]. The theoretical analysis of [1] has shown that neural learning can be improved when the statistical characteristic of the target pattern vectors is modified based on the findings. Also, it is well known that speech signals have a unique distribution characteristic, which is already exploited by compression techniques such as A-law or μ-law in order to improve the performance of PCM. A similar distinctive distribution characteristic of speech signals can also be observed in the acoustic or articulatory domains. Applying our theoretical results, these characteristics will be exploited for the parameter estimation of an articulatory speech synthesiser through improved neural learning.

2. IMPROVING NEURAL NETWORK MAPPING

To train a neural network, a training set was created using a simplified, Kelly-Lochbaum vocal tract (VT) model. The assumptions made to simplify the VT implementation are that the VT consists of lossless uniform, concatenated acoustic tubes, that the VT consists of rigid walls and that planar wave propagation is valid. Linear interpolation was applied to the optimised area function of 10 English vowels [5] giving a total of 164 training patterns. Acoustic input pattern vectors $\mathbf{x_i}$ were derived from the transfer function of the VT which has been simulated using the MATCAD software package. The radiation load was approximated by a 1st order IIR filter, setting the reflection coefficient at the boundary of the last section as 0.99 to ensure IIR filter stability.

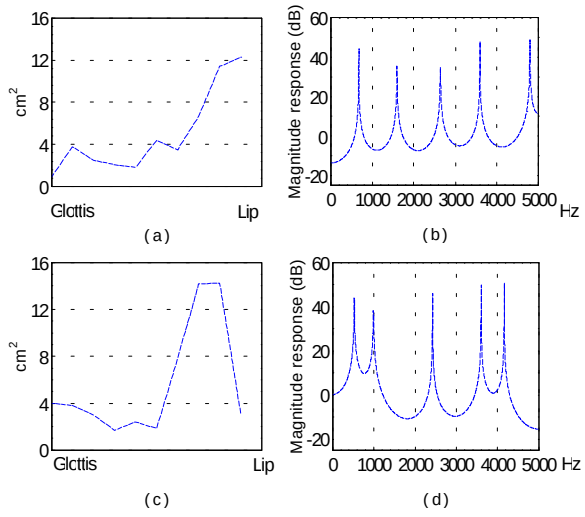

Figure 1 The VT area function and corresponding transfer functions: (a) and (b) for vowel /ae/ , (c) and (d) for vowel /ao/

Two examples of training set vectors are shown in Figure 1 and corresponding articulatory and acoustic vectors are given in Table 1a and Table 1b. for vowel /ae/ and /ao/.

	Area function (cm^2)									
/ae/	0.99	3.42	2.42	2	1.74	4.77	3.25	6.18	12.07	13.04
/ao/	4	3.8	3	1.7	2.4	1.85	7.8	14.2	14.25	3.1

Table 1a. Articulatory vector values for /ae/ and /ao/

	Formant Frequencies (Hz)				
/ae/	686	1,608	2,611	3,541	4,776
/ao/	526	991	2,424	3,604	4,163

Table 1b. Corresponding acoustic vector values

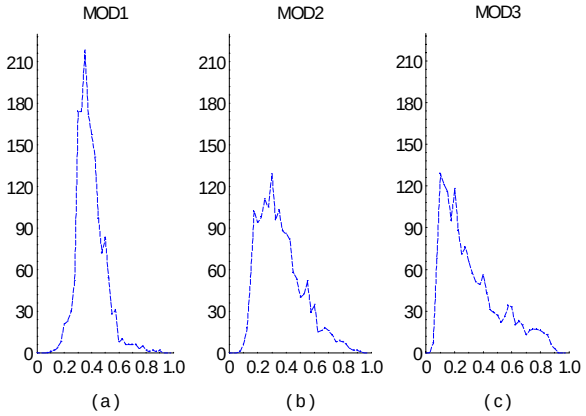

Figure 2. The distribution characteristic of the three training sets modified by MOD1, MOD2 and MOD3

The articulatory target pattern vectors t_o , which are also known as the area function, are subject to three modifications MOD1, MOD2 and MOD3. Figure 2 shows the distribution characteristic of the articulatory target patterns after the modifications. Also Table 2 gives the statistical characteristic of the modified articulatory target pattern vectors t_o in terms of expected value, standard deviation and their average Euclidean distance after each modification.

Type of modification	Expected Value	Standard Deviation	Average Euclidean Distance
MOD1	0.4049	0.102	0.0339
MOD2	0.375	0.1552	0.0524
MOD3	0.3279	0.2034	0.0672

Table 2. The statistical characteristics and average Euclidean distance of the articulatory target vectors after each modification.

Two different NN structures, with 15 and 18 hidden nodes, have been trained in order to see the effect of pre-processing the articulatory target pattern vectors t_o on learning and generalisation. The learning rate η and momentum term α have been selected as 0.01, and 0.3 for the NN with 15 hidden nodes, and as 0.1 and 0.2 for the NN with 18 hidden nodes. Each NN has been trained to reach the error threshold, which is 0.002 mse, otherwise the learning process is suspended at the iteration threshold of 10^5. The same initial starting point has been used in all cases. The acoustic input pattern vectors are also modified as stated in [6]. Learning curves for the NN having a 5-15-10 structure are shown in Figure 3. The results are also given in Table 3 and show the acceleration rate in convergence and the number of iterations that each NN has required.

Type of articulatory vector modification	NN structure 5-15-10		NN structure 5-18-10	
	Raw Acoustic Vectors	Modified Acoustic Vectors	Raw Acoustic Vectors	Modified Acoustic Vectors
MOD1	29,500	3,950	9,850	800
MOD2	100,000	23,900	45,650	5,200
MOD3	100,000	100,000	100,000	68,650

Table 3 The number of iteration required to train a NN

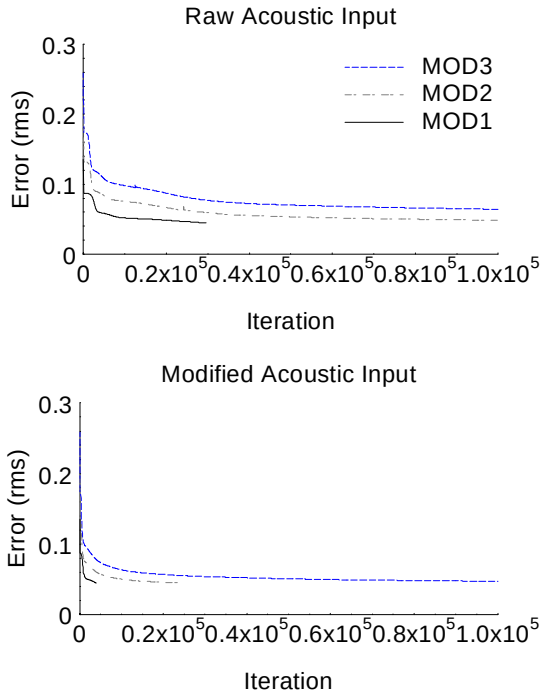

Figure 3. The error curves for the NN having a 5-15-10 structure trained with the raw acoustic input (above), and modified acoustic input (below). Articulatory target vectors are subject to three different preprocessings, namely MOD1, MOD2 and MOD3, in both cases.

The acceleration rates for the NN with the 5-18-10 structure are calculated, when modified acoustic input is employed, as a factor of 85.81 for the case of MOD1 and a

factor of 13.20 for the case of MOD2 when compared to the convergence rate in the case of the MOD3 modification. The results are in line with the theoretical expectations drawn in [1], showing that the modification of the statistical characteristic of the articulatory target vectors t_o has a great effect on the speed of neural learning.

3. ESTIMATION PERFORMANCE OF THE NN's

The estimation performance of all attempts has been calculated using a set of acoustic-to-articulatory patterns that had not been used in the learning phase. After the training process, each trained network was presented with the test pattern set. The estimated area functions are subject to post-processing in order to retrieve the actual values of the area functions. In order to correctly evaluate the performance of each NN, the error between the original and estimated area functions obtained by each of the NNs is calculated as a RMS error. Then, based on the estimated area function a vocal tract configuration is used to reconstruct the corresponding impulse response of the vocal tract using the simulated Kelly-Lochbaum model in MATCAD. The error between the original impulse spectra and the reconstructed impulse spectra is calculated as a RMS error.

The acoustic-to-articulatory inverse problem is an ill posed one due to the absence of a one-to-one mapping and due to the insensitivity of the set of parameters, i.e. the articulatory or acoustic parameters, to the changes in the other. The impulse spectra of the vocal tract does not show any significant changes when a big disturbance in the area function of the vocal tract takes place, as happens in some cases, while small changes in the area function could result in an enormous difference in the corresponding impulse spectra. These factors mean that evaluation of the NN performance can be achieved through studying the sum of the RMS errors. For an illustration of these problems, a set of results for vowels /IH/ are given in Table 4a and Table 4b.. The results show that despite the increase in the RMS error of the estimated articulatory parameters, from 1.4499 to 1.6065, in the case of MOD1, the RMS error of the reconstructed impulse spectra is reduced from 9.5442 to 7.9460, as shown in column 3 of Table 4b. Also, the obtained RMS error for the reconstructed spectra is the smallest value, as shown in Table 4b, when modified acoustic input vectors and articulatory modification of MOD2 is employed, contradicting the higher RMS error of the estimated articulatory parameters as seen in Table 4a.

Thus, the performance of the NN is measured against the total RMS error. Based on this performance measure, results given in Table 5 show that despite the speeding-up effect on learning, in the case of MOD1, the NN exhibits poor generalisation as expected and the reason for this has already been explained in the previous section.

Vowel /IH/	Type of Articulatory Modification		
	MOD3	MOD2	MOD1
Raw acoustic	0.8389	1.2541	1.4499
Modified acoustic	0.6201	1.1674	1.6065

Table 4a. The RMS error of the estimated articulatory parameters

Vowel /IH/	Type of Articulatory Modification		
	MOD3	MOD2	MOD3
Raw acoustic	8.2955	9.5442	10.4383
Modified Acoustic	8.019	7.946	9.9416

Table 4b. The RMS error of the reconstructed acoustic parameters

It is also worth noticing that the total RMS error has a linear relationship with respect to the average Euclidean distance. Incrementing the average Euclidean distance implies a decrease in the total RMS error, as can be seen from Table 2 and Table 5.

Type of modification	RMS error in Area Function	RMS error in Impulse Spectra	Total RMS error
MOD1	1.5024	7.9104	9.4128
MOD2	0.8749	6.9909	7.8658
MOD3	0.8687	6.7081	7.5768

Table 5. The RMS error in the estimated acoustic and articulatory parameters
(NN structure is 5-18-10, η=0.1 α=0.2)

Thus in the next section further modification, which enhances the average Euclidean distance of the target pattern vectors t_o, will be introduced to remedy this problem, at the expense of the convergence speed.

4. RE-OPTIMISING THE PREPROCESSING PROCEDURE IN FAVOUR OF IMPROVED GENERALISATION: MOD4

In order to obtain a good estimation performance, whilst retaining the enhanced convergence speed, the articulatory target pattern vectors modified according to MOD1 are subject to further pre-processing which results in a higher average Euclidean distance within the target vectors. This modification is referred to as MOD4 and is given as follows

$$(1)\ A_{re-optimised} = \frac{\alpha + \beta}{\alpha + \exp\left(\delta - \omega \times A_{opt}\right)}$$

where $\alpha = 2.0$, β=–0.1, δ=5.0, ω=10.0 and A_{opt} is the articulatory target vector after MOD1.

The statistical characteristics of this further processed articulatory target vector are given in Table 6. for comparison purposes. As can be seen the average Euclidean distance for this training set is increased from 0.0339 to 0.0654. Despite the higher standard deviation which has an adverse effect on neural learning according to the finding of the analytical analysis, the

expected value m_{t_o} is increased from 0.4049 to 0.4149 which is consistent with the requirement of the optimum point for x_o stated [1].

Type of modification	Expected Value	Standard Deviation	Average Euclidean
MOD1	0.4049	0.102	0.0339
MOD3	0.3279	0.2034	0.0672
MOD4	0.4149	0.1907	0.0654

Table 6 Statistical values of the pre-processed articulatory output vectors with optimum pre-processed acoustic input vectors

The NN with 18 hidden nodes has converged to the predefined error threshold within 22,500 iterations. Despite, as expected, that modification MOD4 reduced the convergence rate, when compared to MOD1, it is still quicker than when MOD3 was applied. Also the convergence rate of the NN is considerably faster when compared to MOD2 in the case of training with the raw acoustic input vector.

Type of modification	RMS error in Area Function	RMS error in Impulse Spectra	Total RMS error	Change in RMS error as percent
MOD1	1.5024	7.9104	9.4128	+19.67%
MOD2	0.8749	6.9909	7.8658	-----
MOD3	0.8687	6.7081	7.5768	-3.67%
MOD4	0.9577	5.7937	6.7514	-14.67%

Table 7. A comparison of RMS error for a NN with pre-processed articulatory target vectors

After the training process, the NN has been tested with the same test pattern set. The results show that the total RMS error is reduced to 6.7514 from 9.4128 which is a 28.27% enhancement.

As a result, the solution to the inverse problem for the acoustic-to-articulatory mapping is improved through applying a further optimisation of the articulatory target vectors, at the expense of convergence speed.

5. CONCLUSIONS

Using an acoustic-to-articulatory mapping problem, the theoretical study of neural learning in [1] has been validated through improving the solution to the inverse problem in articulatory speech synthesis. Through altering the statistical characteristics of the target pattern vectors the total RMS error has been reduced and acceleration in learning has been obtained, when compared to training using raw acoustic and articulatory pattern vectors.

REFERENCES

[1] Altun, H. and Curtis K. M. 1998. Evaluation of neural learning in a MLP NN under different statistical characteristics of target pattern vectors, (Submitted to Neural Processing Letters)

[2] Klatt D. H. , 1987. Review of Text-to-Speech Conversation for English, *The Journal of Acoustical Society of America* ,vol.82, pp. 737-793.

[3] Schroeter J. and Sondhi M. M. 1994. Techniques for Estimating Vocal-Tract Shapes from the Speech Signal, *IEEE Transactions On Speech And Audio Processing*, vol.2, pp. 133-149.

[4] Fant G. 1991., What can basic research contribute to speech synthesis. *Journal of Phonetics*, **19**, pp 75-90

[5] Rahim M. 1991. Neural Networks in Articulatory Speech Synthesis, PhD thesis, University of Liverpool, UK.

[6] Altun, H. and Curtis, K.M.. 1998. Exploiting the statistical characteristic of the speech signal for improved neural learning in a MLP neural network, *IEEE Neural Networks for Signal Processing VIII*, Edited by Tony Constantinides et al, ISBN 0-7803-5060-X, pp 547-556.