VOCAL TRACT SHAPE TRAJECTORY ESTIMATION USING MLP ANALYSIS-BY-SYNTHESIS

Hywel B. Richards† *John S. Bridle*‡ *Melvyn J. Hunt*‡ *John S. Mason*†

†Department of Electrical & Electronic Engineering, University of Wales Swansea, SWANSEA, SA2 8PP, UK.

‡Dragon Systems UK Ltd, Millbank, Stoke Road, Bishops Cleeve, CHELTENHAM, GL52 4RW, UK.

email: h.b.richards@swansea.ac.uk

ABSTRACT

The objective of this work is a computationally efficient method for inferring vocal tract shape trajectories from acoustic speech signals. We use an MLP to model the vocal tract shape-to-acoustics mapping, then in an analysis-by-synthesis approach, optimise an objective function that includes both the accuracy of the spectrum approximation and the credibility of the vocal tract dynamics.

This optimisation carries out gradient descent using back-propagation of derivatives through the MLP. Employing a series of MLPs of increasing order avoids getting trapped in local optima caused by the many-to-one mapping between vocal tract shapes and acoustics. We obtain two orders of magnitude speed increase compared with our previous methods using codebooks and direct optimisation of a synthesiser.

1. INTRODUCTION

Articulatory representations offer potential benefits over usual spectral representations. The physical limitations of the dynamics of the articulators in the vocal tract imply slowly changing parameters. A closer relationship with the phonetic domain suggests that such representations might be more suitable for recognition, and, as it is believed that coarticulation and transitional effects take place at this level, offer hope in modelling these phenomena.

We focus here on physical representations in the form of area functions rather than the relative positions of the lips, tongue, jaw, etc. Our motive is to achieve a representation exhibiting the above desirable qualities, rather than estimate the precise geometry of the vocal tract.

The problem of estimating vocal tract shapes from the speech signal, often termed the inversion task, is difficult because this mapping is both non-linear and one-to-many. As the mapping is non-linear previous approaches have used techniques such as articulatory codebooks [1], analysis-by-synthesis [2], and a number of continuous non-linear mapping techniques [3].

Previous attempts at using neural networks for the inversion task have generally used multi-layer perceptrons (MLPs) directly, estimating vocal tract shapes from spectral inputs. To overcome the one-to-many mapping problem, one approach is to use multiple MLPs, each mapping a region of articulatory space, with the appropriate MLP being selected by a final dynamic programming search of the possible outputs [4]. Alternatively, a sequence of frames can be presented to an MLP to incorporate context to alleviate the uncertainty [5].

In contrast, our approach uses an MLP to synthesise spectra from vocal tract shapes, and this avoids one-to-many mappings in the training of the MLP. This MLP is then used in an analysis-by-synthesis procedure (Figure 1). The MLP is computationally efficient and provides a convenient means of obtaining derivatives for the overall system optimisation. However, as with other analysis-by-synthesis schemes, this method is vulnerable to local optima in the search space, which we address with a hierarchy of MLPs (see Section 6).

In this paper we first describe an analysis-by-synthesis scheme which allows the use of a general n^{th} order model for the articulatory dynamics, and then how the articulatory synthesiser in this scheme can be replaced by an MLP. Finally, we address the problem of initialising this system, in order to reduce its vulnerability to local optima.

2. ANALYSIS-BY-SYNTHESIS

To estimate a vocal tract shape sequence, $\underline{A}(t)$, a cost function, C, is minimised which consists of two components: the acoustic difference between the observed speech and that synthesised from $\underline{A}(t)$, and a continuity cost on $\underline{A}(t)$ (Equation 1). These two components are combined using a weighting factor, k, which applies an appropriate scaling to the two costs [1].

$$C = k \sum_{t=0}^{T-1} \left| \underline{f}\left(\underline{A}_t\right) - \underline{c}_{ot} \right|^2 + \sum_{t=2}^{T-1} \left| \underline{A}_t + a_1 \underline{A}_{t-1} + a_2 \underline{A}_{t-2} \right|^2 \quad (1)$$

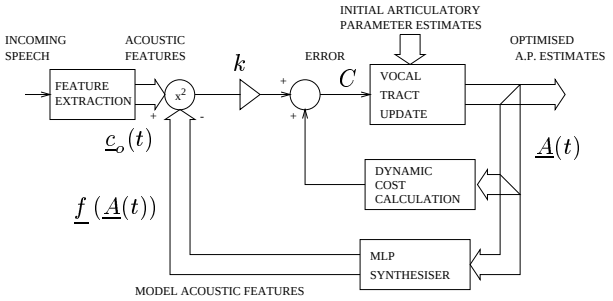

Figure 1: Using an MLP in an analysis-by-synthesis scheme with the inclusion of dynamic constraints.

The first component gives the acoustic cost: the difference between the observed speech spectral vector, $\underline{c}_{ot}$, and the synthesiser spectral output vector, $\underline{f}(\underline{A}_t)$, given the proposed vocal tract shape vector at time t, $\underline{A}_t$.

The second component, the dynamic or continuity cost, gives the deviation of the vocal tract time sequence, $\underline{A}(t)$, from some linear dynamic model, in this case a general second order model $H(z) = 1/1 + a_1 z^{-1} + a_2 z^{-2}$.

We have described an approach to finding a sequence of vocal tract shapes that minimises C using a codebook of shapes, and Dynamic-Programming [1]. It is limited in practice to first-order continuity costs, produces quantised results, and is quite expensive.

The gradient of C with respect to the articulatory parameters is given by

$$\frac{\partial C}{\partial \underline{A}_t} = 2k.\,\mathbf{J}_f(\underline{A}_t)^T.\left(\underline{f}(\underline{A}_t) - \underline{c}_{ot}\right) + 2\left(\left(1 + a_1^2 + a_2^2\right)\underline{A}_t + a_1(a_2+1)\left(\underline{A}_{t-1} + \underline{A}_{t+1}\right) + a_2\left(\underline{A}_{t-2} + \underline{A}_{t+2}\right)\right) \quad (2)$$

where $\mathbf{J}_f(\underline{A}_t)$ is the Jacobian matrix of function $\underline{f}(\underline{A})$ for a given vocal tract shape $\underline{A}_t$, the elements given by

$$\mathbf{J}_f(\underline{A}_t)_{(i,j)} = \frac{\partial f_i(\underline{A}_t)}{\partial A_{jt}} \quad (3)$$

We have experimented with a successful but computationally expensive scheme in which an articulatory synthesiser is used directly in the analysis-by-synthesis loop [6]. The Jacobian was estimated by perturbing along each axis of the articulatory space A_{jt}.

In contrast the MLP is less expensive to use as a synthesiser, and in particular the required derivatives can be computed by back-propagation.

The training of such an MLP, and its subsequent use in our analysis-by-synthesis system to estimate vocal tract shapes is discussed in the next sections.

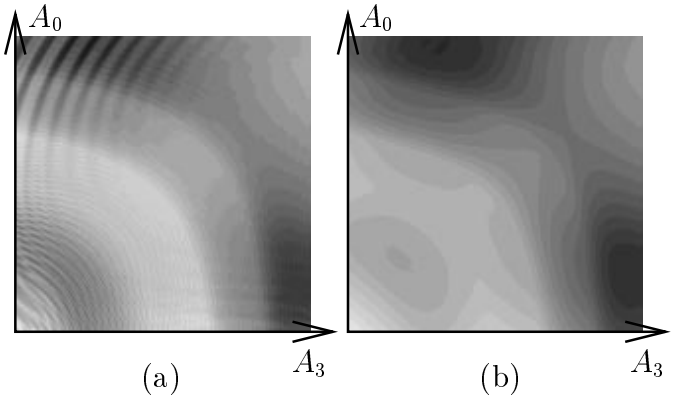

Figure 2: The acoustic cost function for an [r] spectrum with two articulatory parameters (a) without and (b) with distributed losses taken into account.

3. ARTICULATORY-ACOUSTIC MAPPING

We have used 10^{th} order RPS-weighted PLP cepstral coefficients [7] as the acoustic representation, $\underline{c}_{ot}$, and a four-parameter constrained version of the Distinctive Regions Model [1] [8] as the articulatory representation, $\underline{A}_t$.

Our studies of the articulatory-acoustic mapping have shown that the inclusion of realistic distributed losses into the vocal tract model introduces a useful stability to the conversion between the two domains [6]. We illustrate further this characteristic in Figure 2, which shows the acoustic cost as a function of two articulatory parameters, for a particular target spectrum. As the whole tract shape is here defined by four articulatory parameters, the figure shows a two-dimensional slice through a four-dimensional articulatory space. Darker regions correspond to areas of the space where the acoustic output of the model is closer to that of the observed speech, in this case an [r] sound.

It can be seen from Figure 2 that the use of losses, besides making the mapping more realistic, results in a smoother acoustic cost function, implying a smoother articulatory-to-acoustic mapping. This is important for iterative gradient descent techniques, which perform better on such smooth error curves, and also for the success of an attempt to approximate the mapping with an MLP.

Figure 2 also exhibits an example of the many-to-one mapping, as two distinct regions of the articulatory space give a good fit to the observed acoustics. Closer examination of this function in all four dimensions reveals that these regions do not constitute a distinct bimodality, but are in fact connected in a banana-like shape. Despite this connection, such a complex region of low acoustic cost could conceivably 'capture' (in a sub-optimal solution) an iterative search which simultaneously tries to minimise a continuity cost.

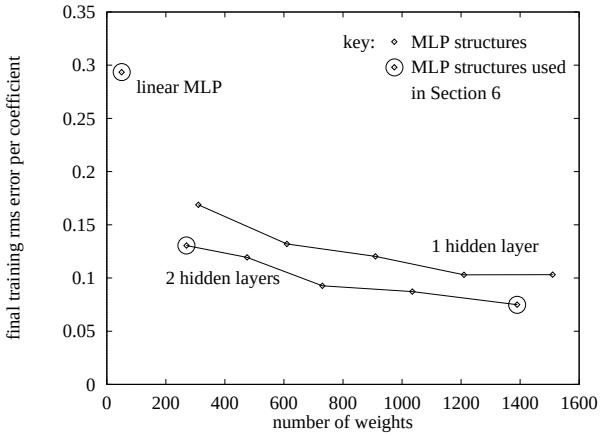

Figure 3: Final training error for different MLP topologies. MLPs with 10 to 30 units in each hidden layer are shown. The MLPs used in Section 6 are highlighted. For reference, the rms 'error' with respect to the mean of the training data is 0.38.

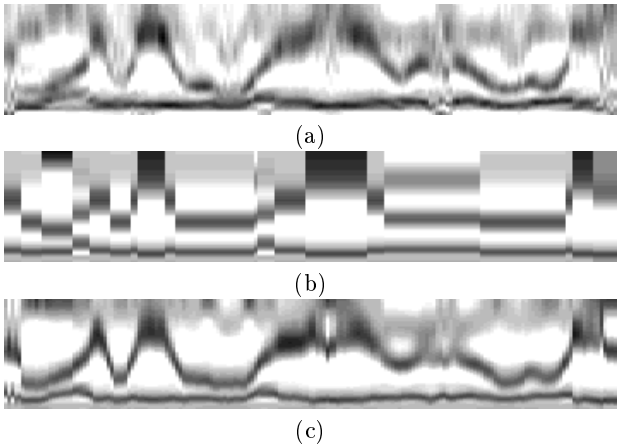

Figure 4: MLP analysis-by-synthesis for the utterance 'Why were you away a year Roy'. PLP-derived spectrograms are shown of (a) the original speech, (b) the spectral output of the vocal tract shapes used for initialisation from a very small articulatory codebook (81 entries), and (c) the spectral output of the optimised vocal tract shapes.

4. MLP TRAINING

Various sizes of MLP from a simple linear net to a two hidden-layer net with 30 units in each hidden layer were trained with the mapping, supplied in the form of an articulatory codebook with 50625 elements. 10% of the training examples were reserved for cross-validation, the error of which failed to turn upwards with continued training in each case. This, together with the profile in Figure 3, would seem to justify the use of an even larger net.

Figure 3 shows the final training error of the MLP with respect to the training data against the MLP size. It can be seen that MLPs with two hidden layers consistently give a more faithful mapping for the same number of weights than those with one hidden layer, suggesting that the mapping is better approximated by a smaller number of higher-order functions. The final training error of the linear net is also shown (top left) for comparison.

5. TESTING AND INITIALISATION

After training the MLP, it can be used in the analysis-by-synthesis scheme of Figure 1. For a 2-hidden-layer net with 30 units in each hidden layer, the vocal tract shape estimates from this scheme gave a superior acoustic fit than quantised codebook estimates. This is despite the fact that this same codebook is used for the MLP training, and shows that the MLP generalises well between the training points.

To give a fair assessment of the shape estimates $\underline{A}(t)$, the original articulatory synthesiser, as opposed to the MLP approximation, should be used in the calculation of the acoustic fit. For example, when the above experiment was repeated for some of the smaller MLPs the acoustic error *increased*. This was despite the fact that the optimiser error calculated using the MLP was decreasing, showing the disparity between the actual mapping and the (in this case oversimplified) MLP approximation.

Since we are using a local, gradient-based search technique with non-linearities, it is not surprising that it is possible for the process to become stuck in local minima of C. Initialization with a static uniform tube is not adequate, but in experiments with the use of the codebook-based method [1] to initialise the search, it was found that a codebook with only 81 entries was adequate to avoid local minima for the example in Figure 4.

6. LINEAR MLP INITIALISATION

Smoother mappings, which yield smoother error functions, have the advantage that local minima in the error curve are less frequent. An approach to avoiding such local minima in gradient descent is to first find a minimum of a smooth approximation of the given error curve, and use this solution to provide the initialisation of the gradient descent of the original unsmoothed error curve. More generally, we can gradually increase the complexity of the mapping during the estimation procedure, starting with a smooth mapping which gives few local minima, and then, as $\underline{A}(t)$ converges towards some approximate solution, the mapping can be made less smooth and more accurate, to give a more accurate solution.

Simpler (or in other words, smaller) MLPs provide smoother mappings. This has been observed from the acoustic cost functions that differently-sized MLPs yield. In the special case of a linear mapping (no hidden layers), no secondary minima can occur as the error function is quadratic from Equation 1 (despite the fact that linear mappings can be many-to-one).

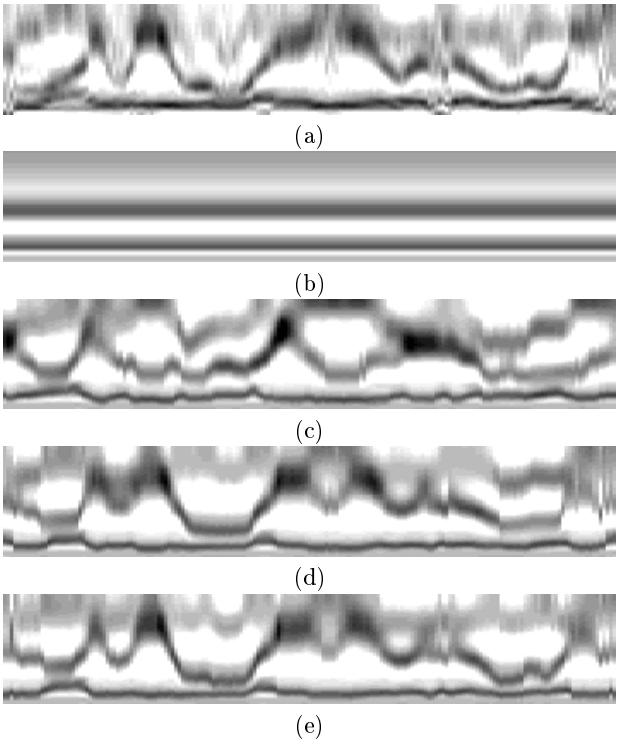

Figure 5: Linear MLP initialisation of MLP analysis-by-synthesis of utterance 'Why were you away a year ago Roy'. PLP-derived spectrograms are shown of (a) the original speech, and the spectral output of the vocal tract shapes of (b) the uniform tube initialisation, and after optimisation using (c) the linear MLP, (d) a 2 hidden-layer MLP with 10 hidden units in each layer, and finally (e) a 2 hidden-layer MLP with 30 hidden units in each layer.

Such an approach has proved effective as an initialisation procedure for the MLP analysis-by-synthesis. Starting with a linear net, and gradually increasing the size of the MLP used, gradient descent is carried out on the resulting cost function of Equation 1, each time the solution used to initialise the gradient descent of the successive, more accurate cost function. In Figure 1, this is simply changing the articulatory synthesiser section during the optimisation while keeping $\underline{A}(t)$, and all other things the same. This is very similar to the Graduated Nonconvexity algorithm described by Blake [9].

Figure 5 shows an example of applying this approach to the MLP analysis-by-synthesis, where a hierarchy of MLPs of increasing complexity have been used to obtain accurate articulatory estimates while avoiding local minima. The MLP is changed twice: from a linear net to a two hidden-layer net with 10, and then 30 units in each hidden layer.

An important question that must be answered is: does this satisfactorily resolve the one-to-many problem? Intuitively, a smoothed error function where a bimodality previously existed would be a shallower error curve whose minima would lie somewhere between the two original minima. The dynamic contribution to the cost function would dominate such a shallow curve, and so the dynamic costs here provide some alleviation of the one-to-many problem as desired by the inclusion of such a constraint.

7. CONCLUSIONS

An MLP has been successfully used as a synthesiser in an iterative analysis-by-synthesis technique, significantly reducing the computational effort in estimating vocal tract shapes from the speech signal. For successful training of this MLP, it is beneficial to incorporate losses into the articulatory model used to provide the training examples. This serves not only to improve the synthesis realism, but also to smooth the mapping: necessary for the success of both the MLP approximation and any subsequent gradient descent techniques.

It was found that the analysis-by-synthesis technique is sensitive to its initialisation, and an initialisation procedure using a linear net has been developed in order to avoid local minima in the solution space.

8. REFERENCES

1. H. B. Richards, J. S. Mason, M. J. Hunt, and J. S. Bridle. Deriving articulatory representations of speech. In *Proc. Eurospeech-95*, pages 761–764, 1995.
2. J. Schroeter, J. N. Larar, and M. M. Sondhi. Speech parameter estimation using a vocal tract/cord model. In *Proc. ICASSP-87*, volume 1, pages 308–311, 1987.
3. J. Schroeter and M. M. Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Trans. Speech and Audio Processing*, 2(1):133–150, January 1994.
4. M. G. Rahim, C. C. Goodyear, W. B. Kleijn, J. Schroeter, and M. M. Sondhi. On the use of neural networks in articulatory speech synthesis. *J. Acoust. Soc. Am.*, 93(2):1109–1121, 1993.
5. G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. *J. Acoust. Soc. Am.*, 92(2):688–700, 1992.
6. H. B. Richards, J. S. Mason, M. J. Hunt, and J. S. Bridle. Deriving articulatory representations from speech with various excitation modes. In *Proc. ICSLP-96*, pages 1233–1236, 1996.
7. H. Hermansky, B. A. Hanson, and H. Wakita. Perceptually-based linear predictive analysis of speech. In *Proc. ICASSP-85*, volume 1, pages 509–512, March 1985.
8. M. Mrayati, R. Carré, and B. Guérin. Distinctive regions and modes: a new theory in speech production. *Speech Communication 7*, pages 257–286, April 1988.
9. A. Blake. Comparison of the efficiency of deterministic and stochastic algorithms for visual reconstruction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11(1):2–12, 1989.