DYNAMIC CONSTRAINT WEIGHTING IN THE CONTEXT OF ARTICULATORY PARAMETER ESTIMATION

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

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

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

email: hywel@dragon.co.uk

ABSTRACT

This paper describes a cross-validation method to determine the appropriate weight with which dynamic constraints should be applied when estimating vocal tract shapes from speech. This data-dependent method can estimate the weighting without the need for separate prior knowledge of the source and noise statistics.

The principles are first demonstrated on a simple one-dimensional system analogous to speech production. As the data here is synthetic, the statistics are known, and so the success of the method can be objectively assessed.

Next, the same principles are extended to real speech to improve the estimation of vocal tract shape trajectories.

1 INTRODUCTION

For speech coding, recognition, and synthesis, it is believed that an articulatory description may be preferable to conventional acoustic features because physical limitations of the vocal tract imply slowly changing parameters. Also, the close relationship between the articulatory and phonetic domains suggests that an articulatory parameter set could be a more appropriate representation for recognition.

Despite such attractions, articulatory parameters are not commonly used because of the difficulty in estimating them from speech. There is a complex mapping between the acoustic and articulatory domains, which is both non-linear and non-unique [1] [2]. Such a relationship requires the use of non-linear mapping techniques such as neural networks, non-linear regression or the use of articulatory codebooks as reviewed in [2].

In previous work we have attempted to overcome the difficulties mentioned by using articulatory codebooks with a dynamic programming search to impose dynamic constraints on the estimated vocal tract shape sequences [3]. Unvoiced sounds pose a particular problem for vocal tract shape estimation, as the observed acoustic signal carries very little information about the vocal tract shape behind the constriction. Previously, we have addressed this problem by extending the dynamic programming method [4].

Also in recent work, we have addressed the problem of computation associated with the dynamic programming approach by developing an alternative method using MLP analysis-by-synthesis [5]. Here, an MLP is used to synthesise speech spectra from a hypothesised vocal tract shape sequence. This is compared with the observed speech, and the hypothesis altered to give an improved match. This cycle is repeated until the hypothesis converges to a solution. The problem of local minima in the solution space was also addressed. These were avoided by optimising for a hierarchy of MLPs of increasing complexity.

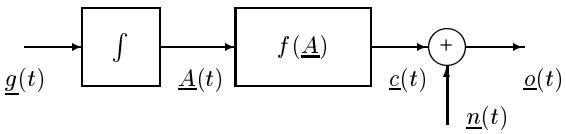

Figure 1: A block diagram of the speech production model. A Gaussian function drives an integrator to produce the articulatory trajectories, which are then mapped through a non-linear system and distorted by additive noise to give the observed acoustic pattern.

The central task in either the articulatory codebook or the MLP analysis-by-synthesis approach is the minimisation of an objective cost function that can be justified with reference to a very simple model of the way that the observed acoustic pattern is produced from a changing vocal tract shape. In Figure 1 the parameter vector $\underline{A}(t)$ of the vocal tract shape at time t executes a random walk (Brownian motion), and the acoustic vector is some non-linear function of $\underline{A}(t)$, modified by observation noise. The objective criterion of Equation 1 can be derived from a model of the form shown in Figure 1 via the log likelihood of the observations given such a model. It is also possible to use higher order models than the integrator in Figure 1 [5].

$$C = k \sum_{t=0}^{T-1} \left| \underline{f}\left(\underline{A}_t\right) - \underline{o}_t \right|^2 + \sum_{t=1}^{T-1} \left| \underline{A}_t - \underline{A}_{t-1} \right|^2 \qquad (1)$$

The two components of this cost function are combined using a weighting factor, k, which applies an appropriate scaling to the two costs. This weighting is required because the units of the two components of Equation 1 are different. It can be shown that the best value for k is equal to the ratio of the variances of $g(t)$ and $n(t)$ in the model of Figure 1. Previously, we have estimated a suitable value for this constant by making estimates of these two variances. This has involved making assumptions about the rate of articulator movement, and comparing this with average cepstral distances. Other authors have employed a more heuristic approach, for example Schroeter *et al.* [2] set the weighting according to the change in the acoustic parameters, whereas Yehia *et al.* [6] based their weighting upon the certainty of the acoustic measurements.

The purpose of this paper is to describe a data-dependent method which estimates a value for k based solely upon the acoustic signal.

2 ESTIMATING THE COST WEIGHTING

As we have discussed in previous work [3], different values of k are appropriate for different speech sounds and levels of background noise. For example, during quiet periods of speech, such as during a stop closure, the effective variance of $n(t)$ in Figure 1 increases. This is due to the logarithmic compression employed in most acoustic representations, which makes them much more sensitive to additive noise at low signal powers. Consequently, k should be lowered in this case to weight less heavily the now unreliable acoustics.

In this paper we consider just one case, namely vocalic sounds of a fixed signal-to-noise ratio, but the principles are readily extendable to a time-varying k, which could be determined by the speech class and the prevailing signal-to-noise ratio.

The observation noise, $n(t)$ in Figure 1, models the uncertainty in the observed spectrum due to additive noise, etc., but together with $g(t)$, must also absorb the error in the mapping $f(\underline{A})$ and the dynamic model. This is because our approximation to the articulatory-acoustic mapping, and likewise the first-order dynamic model, is not exact for any given speaker or speech segment.

Our data-dependent technique consists of estimating vocal tract shape trajectories for different values of k, using either of the techniques that minimise the cost function in Equation 1, and then comparing the results to see which value of k gives the shape trajectories with the best acoustic fit to the observed data. A glance at the cost function of Equation 1 reveals an obvious flaw in this method: we expect that the largest value of k will always give the best acoustic fit to the observed data, as the largest value of k gives the greatest weight to the acoustic match. For this reason we have applied the principles of *cross-validation* to the estimation of k, which allows us to assess the performance of the estimated trajectories on observed data not used in the optimisation.

To do this, the acoustic data is divided into N sets, one set reserved for validation, and the remaining acoustic information used to estimate the articulatory sequence, $A(t)$, for a given value of k. The acoustic error for the validation set, given this sequence, gives us the validation error. If this is summed over the N sets (selecting each set for validation in turn), and repeated for different values of k, the value which gives the minimum validation error can then be chosen.

3 TESTING THE METHOD

To give an insight into the proposed method, first we apply it to a much simpler, but analogous, problem. A one-dimensional version of the system shown in Figure 1, with a sigmoid nonlinearity used for $f(A)$ and Gaussian driving and noise functions $g(t)$ and $n(t)$, is used to generate a time sequence of observation data $o(t)$ (Figure 2(c)). By generating the data for this problem artificially, the desired sequences $A(t)$ and $c(t)$ are available, and so the performance of the technique can be objectively assessed.

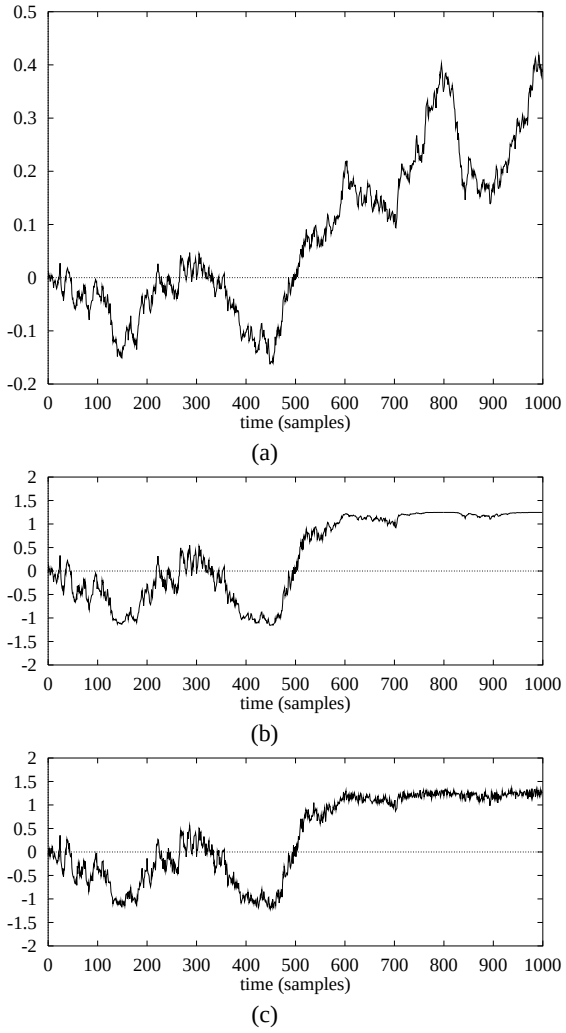

Figure 2: Artificial data for (a) $A(t)$, (b) $c(t)$ and (c) $o(t)$ corresponding to the system of Figure 1.

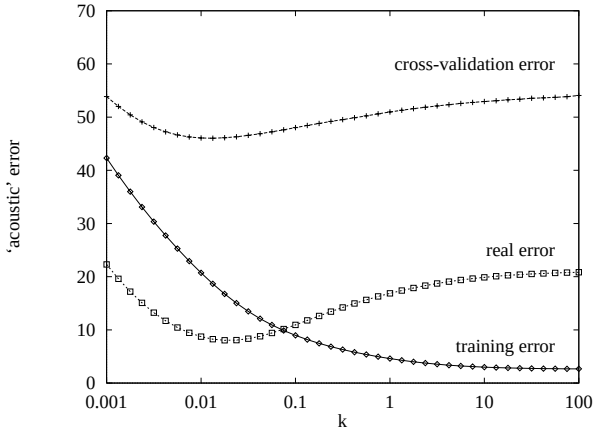

Figure 3: Errors for training and cross-validation data for different values of k when used to analyse the observation sequence $o(t)$ in Figure 2(c). The real error with respect to the undistorted data $c(t)$ is also shown for comparison.

The observation data, $o(t)$, is divided into five sets. This could be done randomly, but as the time-sequence data has a natural order, the members of each set were chosen at equal intervals in the sequence. Each time, a set was excluded, and the remaining data used to estimate the $A(t)$ sequence using the same analysis-by-synthesis method described in [5]. The values of $A(t)$ for the unseen observations in $o(t)$, that is every fifth sample, were interpolated from the estimates for the available data via the inferred vocal tract shape trajectory $A(t)$.

It can be seen from Figure 3 that the 'acoustic' error for the training set decreases monotonically with increasing k, as expected. The cross-validation error also decreases at first, but after a point the error starts to rise again as the underlying dynamic model is neglected and predictions for the unseen data become worse.

Also shown in Figure 3 is the error of the estimated undistorted vocal tract output $c(t)$ with respect to that actually used in the generation of $o(t)$. It can be seen that the minimum of this error curve is close to the position predicted by the cross-validation error.

4 REAL SPEECH

The principles of the previous section have also been applied to vocal tract shape estimation from speech for the two methods that we have previously used: articulatory codebooks with a dynamic programming search [3] and MLP Analysis-by-Synthesis [5].

Training and validation error profiles are shown for the two techniques in Figures 4(a) and (b). The real speech used here was a single utterance of 'Why were you away a year Roy?', and this was tested for a range of k from 0.001 to 100.

As the articulatory codebook method employs a global exhaustive search, then the curves shown in Figure 4(a) are very much as expected. The acoustic error with respect to the 'training' data here decreases monotonically with k as increasingly more emphasis is placed on the acoustic match in Equation 1. The cross-validation error decreases to an unstable minimum at approximately $k = 1$, before rising slowly again with increasing k. It should be possible to stabilise this curve with the addition of more speech data.

The error curves for the MLP analysis-by-synthesis method, however, do not behave in the same way. Figure 4(b) shows that the training error here does in fact have a minimum point at about $k = 0.5$. This means that despite placing an increasing emphasis on the acoustic match in Equation 1, the actual acoustic match achieved degrades. From this we can conclude that the dynamic constraints are instrumental in aiding the iterative search to find the best minimum in the cost function given in Equation 1. When the regulation provided by the dynamic constraints is removed, then the search is more likely to become trapped in a local minimum. The cross-validation error, however, behaves as expected, reaching a minimum at about $k = 0.1$. The curve has a much more definite minimum than that for the articulatory codebook method, perhaps not surprisingly as the minimum point in the training error curve implies a minimum in the cross-validation error curve here.

The different estimated values for k in Figures 4(a) and 4(b) can be explained by differences expected in the variances of $\underline{g}(t)$ and

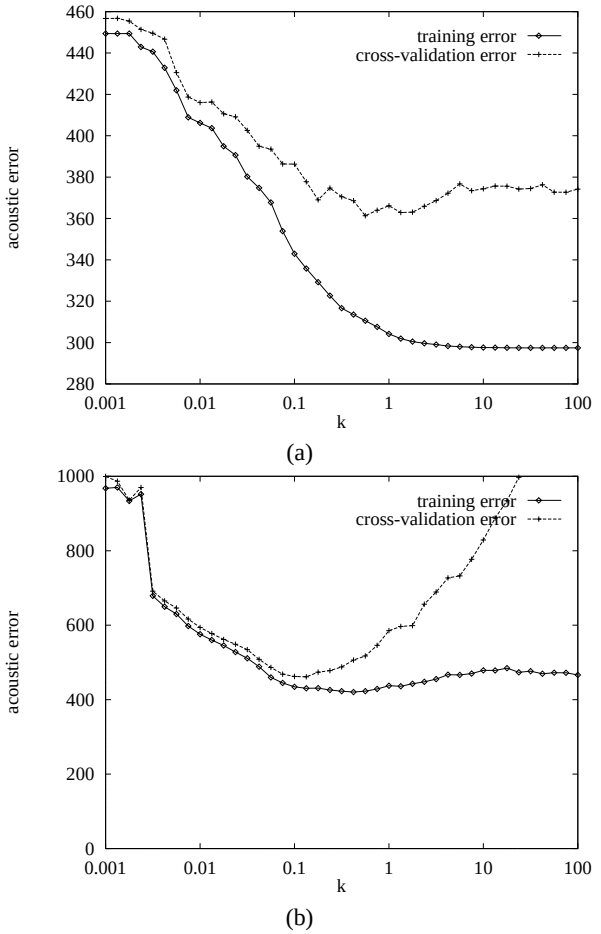

Figure 4: Training and cross-validation error for different values of k for (a) the dynamic programming and (b) MLP analysis-by-synthesis methods.

$\underline{n}(t)$ for the two methods. In the articulatory codebook method, the vector $\underline{A}$ can assume a only fixed number of quantised shapes depending on the codebook size (a 50625 shape codebook is used here). In effect, $\underline{n}(t)$ here models not only the observation noise, but also the quantisation noise introduced by the articulatory codebook. The larger the codebook, the smaller this quantisation noise becomes.

For the continuous MLP analysis-by-synthesis method, however, as the acoustic output is no longer restricted to a discrete number of values, this variance accounts only for the observation noise and the discrepancy between the real articulatory-acoustic mapping and our approximation of it (given by the MLP).

Similarly, the variance of $\underline{g}(t)$ for the dynamic programming method is not only dependent on the articulatory dynamics, but also on the codebook sampling in articulatory space. The higher the resolution of the articulatory sampling, the smaller this variance becomes.

Vocal tract shapes estimated from speech using different values of k are shown in Figure 5. The articulatory codebook method has been used here. In Figure 5(c) ($k = 100$) it can be seen that the parameters change very rapidly with time. This is because dynamic constraints are given little emphasis, and so continuity is sacrificed in order to improve the spectral fit. Figure 5(b) shows the parameter trajectories estimated using a value for k

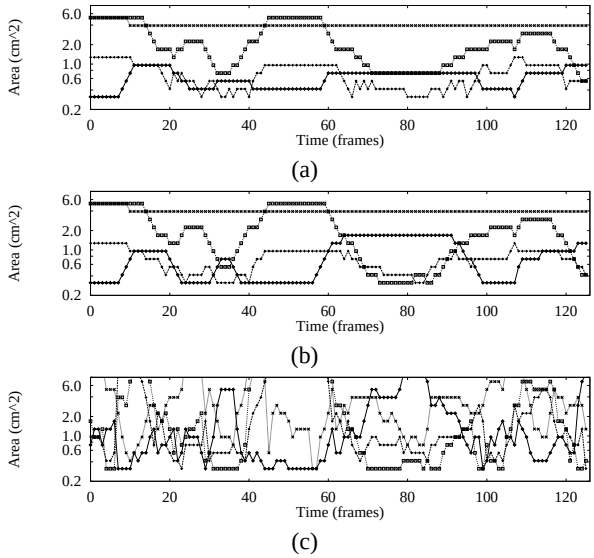

Figure 5: Articulatory codebook derived articulatory parameter trajectories for $k = 0.01$, 1.0 and 100 in (a) to (c) respectively..

very close to the ideal value suggested by Figure 4(a). Here the trajectories are much smoother, bearing in mind the quantisation effect of the articulatory codebook.

The estimated articulatory trajectories for $k = 0.01$ are shown in Figure 5(a). Changes in the articulatory parameters are heavily penalised in the dynamic programming search here. These trajectories are not as flat as might be expected, however, as for computational reasons the search is limited to the n-best acoustic fits from the codebook [3]. For example, using a similar value for k in the MLP analysis-by-synthesis approach yields flat trajectories.

Spectrograms of speech synthesised from these trajectories are shown in Figures 6(b) to (d) respectively. The improved spectral fit for the highest value of k can be seen in Figure 6(d), whereas this fit is noticeably degraded in the 'damped' spectrogram in Figure 6(c), and even more so in Figure 6(b).

The articulatory estimates in Figure 5(b) therefore represent a compromise between spectral fit and articulatory continuity. As we use the ideal value of k here, they also represent the expected observation sequence, given the form of the production model assumed (Figure 1). If the spectrogram in 6(c) does not approximate the original speech in Figure 6(a) very well, it is an indication that this underlying model of speech production could be improved. This is not surprising, considering its simple first-order dynamic component.

5 CONCLUSIONS

A cross-validation method has been proposed to estimate an appropriate value for the weighting of dynamic constraints to yield the best vocal tract shape estimates from speech. This method has the advantage that only acoustic observations are necessary to estimate this weighting.

Artificial data with known statistics has been used to validate the approach, and the method has also been shown to give reasonable results on real data.

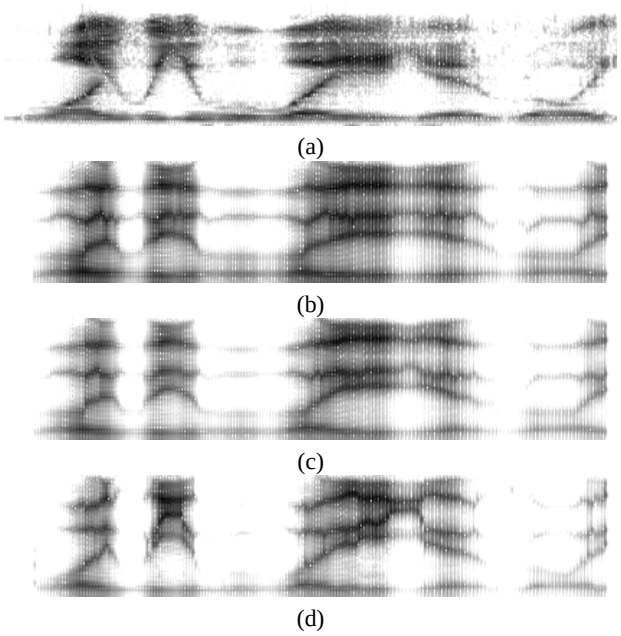

Figure 6: Speech spectrograms synthesised from the vocal tract parameters shown in Figure 5 for $k = 0.01$, 1.0 and 100 in (b) to (d) respectively.. The original speech is shown in (a) for comparison.

However, it is likely that the vocal tract system is far from first order, and the Gaussian assumptions for the driving function and noise are not realistic, partly because the uncertainty they model must also account for the mismatch between the real articulatory-acoustic mapping and our approximation of it, $f(\underline{A})$.

6 REFERENCES

1. B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *J. Acoust. Soc. Am.*, 63:1535–1555, May 1978.
2. 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.
3. 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.
4. 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.
5. H. B. Richards, J. S. Bridle, M. J. Hunt, and J. S. Mason. Vocal tract shape trajectory estimation using MLP analysis-by-synthesis. In *Proc. ICASSP-97*, pages 1287–1290, 1997.
6. H. Yehia and F. Itakura. Determination of human vocal-tract dynamic geometry from formant trajectories using spatial and temporal Fourier analysis. In *Proc. ICASSP-94*, volume 1, pages 477–480, 1994.