DERIVING ARTICULATORY REPRESENTATIONS FROM SPEECH WITH VARIOUS EXCITATION MODES

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

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

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

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

ABSTRACT

A new approach is described which estimates vocal tract shape sequences for speech consisting of voiceless speech and periods of silence as well as voiced speech. This method, based on the use of articulatory codebooks, has proved successful in identifying the place position of stops and fricatives.

Secondly, we focus on voiced speech in particular. A fast analysis-by-synthesis scheme, which gives continuously-valued area estimates, has been developed. Savings in computation of 50:1 have been achieved by using an MLP to perform the synthesis in this method. The technique also allows a more complex dynamic model to be used.

1. INTRODUCTION

Most previous work concerning the estimation of articulatory parameters from the speech waveform considers only the voiced portions of speech. There are good reasons for this restriction: voiced excitation gives a well-defined formant structure to the spectrum, which provides a relatively large amount of information regarding the shape of the vocal tract. Also, the position of the source ensures that resonances corresponding to the whole of the vocal tract shape are well illuminated by the excitation. An alternative, but invasive, approach to obtain a precise spectrum is to use externally generated artificial excitation with known characteristics [1] [2].

However, speech consists not only of voiced sounds, but also of silence periods and unvoiced sounds, and as these sounds can contribute significantly to the perception of speech and provide additional evidence as to the state of the vocal tract, it would be wise to incorporate them in the estimation.

Unfortunately the mapping between the articulatory and acoustic domains is not straightforward and can vary abruptly especially when different excitation types are considered. In the production of unvoiced sounds, the location of the source, the decoupling effect of the constriction, and the high losses at the open glottis result in the resonances of the rear cavity being less prominent in the output spectrum [3]. Also, wider bandwidths, associated with greater losses, make this spectrum less well-defined. During silence intervals in the speech the spectrum contains no information regarding the shape of the vocal tract. In this case the measured spectrum adopts that of the background noise or falls against preset limits in the analysis.

This paper describes a new approach to the use of voiced, unvoiced and silence intervals in the estimation of vocal tract time sequences in an extension to our previous work on the analysis of voiced speech [4]. This new approach enables the application of the same underlying dynamic constraints to all three speech categories, despite the more complicated nature of the articulatory-acoustic relation when considering speech sounds in general.

In the second part of this paper we introduce a new fast method of inversion, suitable for voiced speech. This gives continuously valued vocal tract area estimates, and also allows the use of a more general dynamic model for the changes in these areas.

2. A CODEBOOK APPROACH TO INVERSION

In our previous work on the estimation of articulatory parameters [4] an articulatory codebook containing a range of vocal tract shapes (160 000) and their resulting speech spectra was used to represent the articulatory-to-acoustic domain mapping. This codebook was pre-generated using a lossless Kelly-Lochbaum model whose area function was specified according to the Distinctive Regions Model (DRM) [5], producing the appropriate spectral outputs for given vocal tract shapes that were sampled at logarithmic intervals in area. The perceptually-based PLP cepstral coefficients were used with RPS weighting to provide the spectral representation.

A pre-selection procedure forwarded a subset (typically 1000) of these codebook entries for each frame of speech to a dynamic programming (DP) search, corresponding to those vocal tract shapes which yielded the most acoustically similar output to the observed spectrum for that frame of speech. The DP search then selected from the 1000^T vocal tract time sequence possibilities available (where T is the number of frames), the one that minimised a cost function based on the acoustic similarity between the codebook entries and observed spectra, plus the continuity of the vocal tract shape across the speech frames (Equation 1). $k(t)$ here is the relative weighting of the acoustic cost, which was set to a constant for

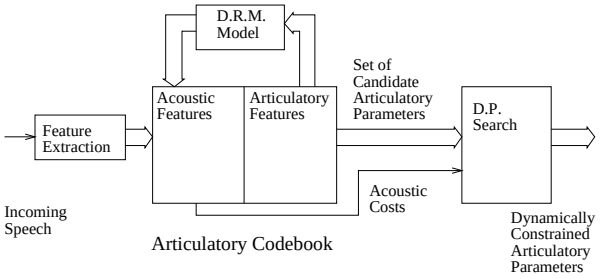

Figure 1: The DP search of an articulatory codebook.

the analysis of voiced speech.

$$C = \sum_{t=0}^{T-1} \left(k(t) \sum_{i=1}^{M} \left(c_{si}(t) - c_{ci}(t) \right)^2 + \sum_{i=1}^{N} \left(A_i(t) - A_i(t-1) \right)^2 \right) \quad (1)$$

This procedure is represented in Figure 1.

3. EXTENSIONS TO THE CODEBOOK APPROACH

The previous analysis has been applied successfully to voiced speech. It would seem possible to apply this technique directly to unvoiced speech by simply adding further codebook entries, but we have found that the characteristics of these sounds make them ill-suited to this approach.

3.1. Unvoiced Sounds

It was found that one of the reasons why the previous method was less successful for unvoiced sounds was that the evidence provided by the spectrum alone was insufficient to make a hard decision at the pre-selection stage. Also, as the vocal tract shape behind the constriction has little influence on the spectrum for unvoiced sounds, then the pre-selection may `saturate' with the same front cavity shape for different rear cavity shape combinations. This effect also highlights a redundancy in repeatedly performing acoustic distance calculations for these shapes, the results of which should be quite similar.

3.2. Using Separate Codebooks

In order to avoid the failure of the pre-selection procedure it was acknowledged that some articulatory parameters are not always specified well by the observed spectrum, in particular those associated with the vocal tract shape behind the constriction for unvoiced sounds, and the whole of the tract during silence intervals. In our modified approach, therefore, it is left to the dynamic constraints alone to reconstruct the parameter values during these times.

An `unvoiced' codebook was generated on the basis of a few simple assumptions:

- to describe unvoiced speech spectra an approximate representation is appropriate, and
- this approximate representation can be based solely upon the front part of the vocal tract, as the shape of the vocal tract in front of the constriction has a dominant effect on the spectrum.

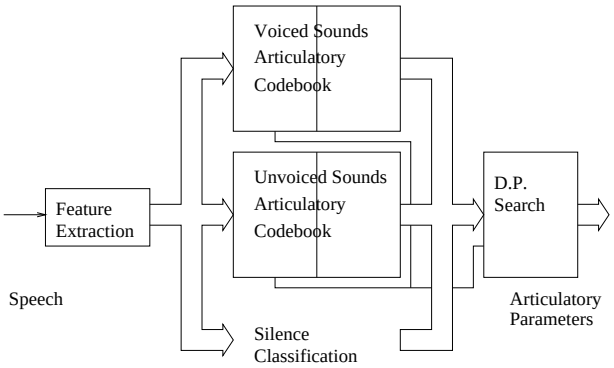

Figure 2: The DP search of possibilities forwarded from two articulatory codebooks after preclassification.

In this way the codebook can be made reasonably small, as only the shape of the front portion of the vocal tract needs to be varied (2364 entries were used), making the technique less sensitive to the pre-selection stage.

For voiced speech, a codebook similar to that used in the previous analysis (Section 2) was used with a reduced size (50625 entries).

3.3. Codebook Selection and DP Search

After an initial classification to label the observed speech as voiced, unvoiced or silent, the analysis considered the spectrum of each speech frame according to its type. An acoustically-based pre-selection search of a voiced or unvoiced articulatory codebook was carried out for voiced or unvoiced speech frames respectively.

A DP search was then made to obtain the best vocal tract time sequence using the previously described matching and dynamic criteria, taking into account that the vocal tract shape for the unvoiced and silence frames is either partially or not at all specified (Figure 2). In the case of the silent frames, given the simple first order dynamic constraints, these frames can be effectively ignored by the analysis and the dynamic cost inversely scaled appropriately according to the length of the silence period. This is equivalent to setting $k(t) = 0$ in Equation 1. A similar approach can be taken for the unspecified parameters for unvoiced sounds, except that additional constraints, such as a constant constriction location, are necessary to ensure the DP search still yields the globally optimum solution.

3.4. Results

The vocal tract shape estimation performed quite well for an initial test set consisting of voiced and unvoiced stops, and fricatives in a VCV context. Place positions consistent with those implied for the given consonants in [g],[p],[t],[k],[f],[s] and [$\int$], with only [b] and [d] failing to yield the correct bilabial and

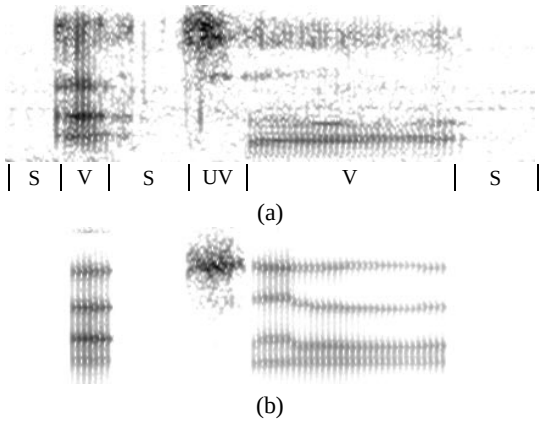

Figure 3: (a) Original and (b) resynthesised speech of an utterance with both voiced (V) and unvoiced (UV) sounds, [ætɑ].

alveolar place positions respectively.

It is felt that in considering the results from these preliminary experiments it would be beneficial to use the additional cue of silence to indicate closure. This is because a complete closure is not always necessary to reproduce the acoustics, and the oversimplified dynamic constraints even discouraged it. A more realistic dynamic model would be an obvious advantage here.

The results also suggest it would be wise to constrain the constriction aperture to sizes consistent with those yielding unvoiced excitation. Constriction widths, most noticeably for [f], were unrealistically large. This can be attributed to the fact that the output spectrum from our model is not sensitive to the constriction size under the second assumption in Section 3.2.

As the current lossless model performs particularly badly (departs from expectation) when small constrictions are used, the inclusion of losses is also deemed to be essential.

4. ANALYSIS-BY-SYNTHESIS

The codebook searching technique allows us a `safe' way of obtaining vocal tract shapes as the search is a global one obtaining the best fit, to the resolution given by codebook quantisation, for the observed speech. If the size of the codebook is sufficiently large then this quantisation effect is minimised. Unfortunately, the DP search described allows only the consideration of first order dynamics for the articulatory parameters. The reasons for this are twofold: the necessity of a Markov process for the DP search [7], and the fact the quantised values from the codebook would provide very poor estimates of higher order derivatives such as acceleration [4].

4.1. Direct Synthesis

An iterative analysis-by-synthesis scheme has been developed which allows the use of an arbitrary n^{th} order model for these articulatory dynamics (Figure 4). This approach is similar to that of Schroeter *et al.* [8], but we have also sought to optimise the dynamic cost simultaneously. Unfortunately, this method incurs large computational costs associated with the repeated synthesis attempts

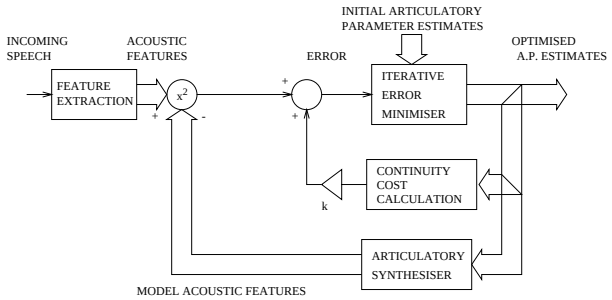

Figure 4: Using the articulatory model for analysis-by-synthesis with the inclusion of dynamic constraints.

and there is also an uncertainty in the result as the iterative search will come to rest in the nearest local cost minima of the solution space.

An investigation of the extent of this local minima problem (which, if only the acoustic cost is considered, is equivalent to the widely documented one-to-many mapping ambiguity between the acoustic and articulatory domains) has been carried out by observing the shape of the acoustic cost function in articulatory space. No distinct bimodalities in this function were encountered, but complex regions of low acoustic cost were observed that could conceivably `capture' an iterative technique which simultaneously tries to minimise a continuity cost in a sub-optimal solution.

Also, it was found that introducing losses in the model stabilises and smoothes the articulatory-acoustic mapping yielding a smoother solution space. This is important for iterative gradient descent techniques which rely on such smooth error curves for their success, and also for the success of an attempt to approximate the mapping with an MLP.

4.2. MLP Mappings

Previous attempts at using neural networks for the inversion task usually use an MLP in the forward direction, evaluating vocal tract shape outputs from spectral inputs. To overcome the one-to-many mapping problem, one approach is to use multiple MLP's, each mapping a region of articulatory space, and the appropriate MLP selected using a final DP search of the possible outputs [9]. Alternatively, a sequence of frames can be presented to an MLP to incorporate context to alleviate this uncertainty [10].

Our approach is to replace the articulatory synthesiser in Figure 4 with an MLP which produces a spectral output, in the form of PLP cepstral coefficients, from vocal tract shape inputs. As the MLP is used in the synthesis direction then the many-to-one nature of the mapping presents no problem to the training. The MLP can be evaluated rapidly, and derivatives of the error function easily back-propagated through the MLP, so this approach represents a considerable saving in computation over using the articulatory synthesiser directly.

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 map-

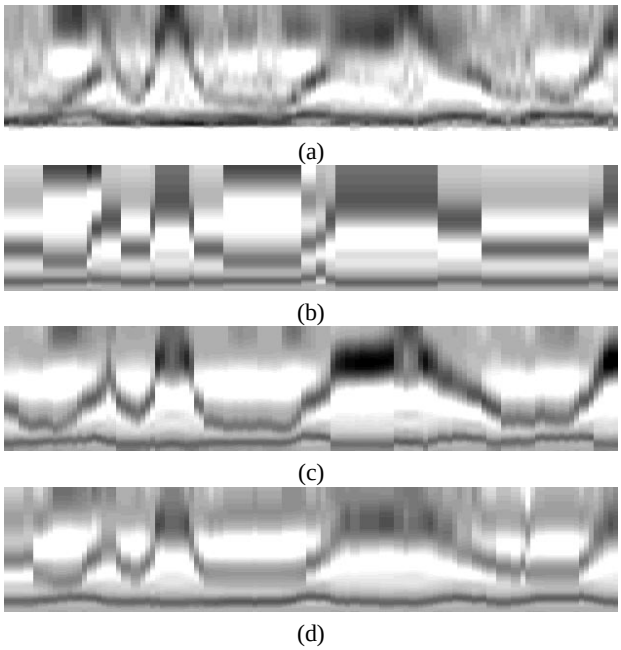

Figure 5: MLP analysis-by-synthesis for the utterance `Why were you away a year Roy'. PLP derived spectrograms are shown of (a) the original speech, (b) the spectral output of the vocal tract shapes used for initialisation from a very small codebook (81 entries), and (c) the spectral output of the optimised vocal tract shapes. (d) shows the equivalent result from a large codebook search (50625 entries).

ping, 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 in each case suggesting that an even larger net might be appropriate for this task.

4.3. Results

The MLP with 2 hidden-layers and 30 hidden units in each hidden layer was used in the iterative analysis-by-synthesis technique shown in Figure 4 for the analysis of `Why were you away a year Roy?' (Figure 5(a)). When initialised with the estimates from a DP articulatory codebook search using the same 50625-element codebook used in the training, the acoustic error (evaluated by direct comparison of original and resynthesised speech) was found to decrease further from that given by the estimates from the codebook search. Replacing the articulatory synthesiser in Figure 4 by this MLP increased the iteration speed by over fifty times.

Together with the DP search of a small codebook, to provide the initialisation (Figure 5(b)), the total inversion task was more than seventy times faster than the large codebook search described in Section 2. A spectrogram of resynthesised speech from vocal tract shapes obtained using this method is compared with those obtained using the codebook approach in Figure 5(c) and (d).

When this experiment was repeated for smaller MLP's the acoustic error was found to *increase*, despite the fact the error calculated by the MLP was decreasing, due to disparity between the actual mapping and the (in this case oversimplified) MLP approximation.

To determine how sensitive the technique is to the initialisation, the optimisation was initialised using the estimates from DP searches using differently sized codebooks. It was found that the technique is indeed sensitive to its initialisation, but a codebook search of a reduced size (and hence a faster process) could be used in the initialisation.

5. CONCLUSIONS

A method has been demonstrated for the estimation of vocal tract shapes across three different classes of speech sound. It was found necessary to treat each of these classes in a different way as the spectrum alone is an unreliable indicator of vocal tract shape for unvoiced sounds and silence. In preliminary experiments this approach appears to be reasonably successful in identifying the correct place position for a range of stops and fricatives, although further enhancements such as the inclusion of distributed losses and the use of silence as a closure cue are recommended.

An MLP has been successfully used as a synthesiser in an iterative analysis-by-synthesis technique, significantly reducing the computational effort. 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 realism of the synthesis, but also to smooth the mapping: necessary for the success of both the MLP approximation and any subsequent gradient descent techniques. The analysis-by-synthesis technique is sensitive to its initialisation, but to provide this it is possible to use a relatively coarsely sampled codebook.

6. REFERENCES

1. H. Yehia, M. Honda, and F. Itakura. Acoustic measurements of the vocal tract area function: sensitivity analysis and experiments. In *Proc. ICASSP-95*, volume 1, pages 652–655, 1995.
2. P. Mermelstein. Determination of the vocal-tract shape from measured formant frequencies. *J. Acoust. Soc. Am.*, 41:1283–1294, May 1967.
3. J. M. Heinz and K. N. Stevens. On the properties of voiceless fricative consonants. *J. Acoust. Soc. Am.*, 33:589–596, 1961.
4. 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.
5. 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.
6. R. Carré and S. Chennoukh. Vowel-consonant-vowel modeling by superposition of consonant closure on vowel-to-vowel gestures. *Journal of Phonetics*, 23:231–241, 1995.
7. J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-time processing of speech signals*. Macmillan, 1993.
8. 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.
9. 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.
10. 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.