ACOUSTIC-PHONETIC MODELLING USING THE HIDDEN DYNAMIC MODEL

Hywel B. Richards and John S. Bridle

Dragon Systems UK R&D, Millbank, Stoke Road, Cheltenham, GL52 4RW, UK.
hywel@dragon.co.uk

ABSTRACT

This paper describes a new approach to acoustic-phonetic modelling, the Hidden Dynamic Model (HDM), which explicitly accounts for coarticulation and transitions between neighbouring phonetic-segments. Inspired by the fact that speech is really produced by an underlying dynamic system, the HDM learns, from labelled speech data, a mapping from a hidden dynamic space where simple dynamic properties exist, to the surface acoustic representation.

In this hidden space, each phone is represented by a single target vector. A simple filter is the dynamic system that interpolates between these targets. A trainable non-linear mapping in the form of a multi-layer perceptron (MLP) maps this hidden dynamic trajectory to an acoustic pattern.

By producing synthetic acoustic patterns using the HDM, we show how it captures the dynamic structure of speech, even with such a economic parameterisation. We also investigate the properties of the learned hidden space, and the effect of varying its dimensionality.

1. INTRODUCTION

Much of the complexity and indirectness of the relationship between the acoustic patterns of speech and the linguistic structures that they represent is caused by context-sensitivity [1].

Conventional large vocabulary continuous automatic speech recognition systems model speech patterns as a sequence of stationary segments (albeit with differential features). Various effects, including phonological variation and hard coarticulation, are dealt with by dividing the contexts of each occurrence of a phoneme into equivalence classes (using context decision trees) and modelling these contextual variants separately (using mixtures of Gaussian components). While this may be a reasonable procedure for some types of phonological variation, it is an extravagant approach to coarticulation, and ignores some well-known properties of real human speech, with the result that very large amounts of training material are required if the system is to perform well with a large vocabulary and a variety of speakers and speaking styles.

The type of system that we have been working towards would model the relationship between phoneme sequences and acoustic patterns in two major steps: phonological rules for the effects that are usually described in terms of segmental reorganisation (e.g. assimilation, elision), and a physically-motivated acoustic-phonetic model that deals naturally with coarticulation (in principle including anticipatory effects across several segments).

In the work reported here we limit ourselves to traditional sequential segmental phonology, and we present a new candidate for just the acoustic-phonetic component, that we call a Hidden Dynamic Model.

2. ALTERNATIVES TO THE HMM

There have been several attempts to find useful alternatives to the frame-by-frame finite-state HMM systems with stationary output distributions within each state. Some approaches treat a whole acoustic segment as a unit, where the segment may be a smaller unit than a phone. In a static segmental model [2] [3] a 'target' for a segment is drawn from a distribution, then each frame has a deviation from this target, as specified by a second distribution. There have been several elaborations in which a non-stationary trajectory (in acoustic space) has parameters drawn from distributions characteristic of the segment label [4] [5].

Another approach is to concentrate on coarticulation between segments, rather than non-stationarity within segments. The classic Haskins work leading to locus theory led to speech synthesis methods in which formant trajectories were constructed using rules of coarticulation [6]. Early attempts to adapt such synthesis models for speech recognition purposes include [7], in which the target formant frequencies and amplitudes were adjusted automatically to optimize a match to formant parameters estimated from natural speech. There have also been attempts to use locus theory directly in HMM systems [8].

In 1991 Raimo Bakis presented a very general model with targets, linear dynamics and non-linear output mapping [9]. The system described in this paper can be seen as a special case of the Bakis model.

Most attempts at dynamic models have used a linear mapping between the space in which the dynamics happens and the acoustic observations [4]. Our approach is similar to that of Blackburn [10] in that we use an MLP, but we use a single MLP and our hidden dynamic system is much simpler (yet quite powerful).

3. THE HDM

The HDM presented here consists of two separate components which together model the relationship between phone sequences and spectra. The first component, driven by a sequence of phones, produces a continuous trajectory in a hidden, 'phonetic' space. The second component, a static non-linear mapping, maps each sample of this hidden trajectory into the corresponding acoustic pattern.

The resulting system (Figure 1) is rather similar to a classic speech-synthesis-by-rule system such as [6] (that converts from phone sequences to formant patterns and thence to waveforms),

except:

- Targets and transitions are more abstract than formant frequencies and amplitudes
- The output is an acoustic pattern (eg smoothed spectrogram)
- Randomness is built in (see below)
- Everything is learnable

To turn this into a stochastic model, to take account of the variability of speech (and the defects of the model), we can introduce random variables in various places. Possibilities include segment timing, per-segment target distributions, per-frame target variations, time constants, and acoustic pattern 'noise'.

The simplest is to treat the whole of the production process as deterministic until the acoustic output is produced, with simple Gaussian 'noise' added to this acoustic pattern to model the variability. This gives us the simple type of HDM described in this paper.

4. THE HDM AS A TRAINABLE SPEECH SYNTHESISER

The Hidden Dynamic Model describes the way in which an acoustic pattern is produced from a sequence of phones with given durations. The structure of the HDM is shown in Figure 1.

For each phone class there is a single target vector which defines a point in the hidden space. For each phone segment in the sequence, the respective target applies for the duration of that segment, resulting in the target sequence, t_j, shown in Figure 1 (j is the time index). This is typically multidimensional, but a single dimension is shown here for simplicity.

Note that the symbols we call 'phones' here are supposed to correspond to acoustic segments with one target – diphthongs and some sets of allophones need more than one target.

The target sequence is smoothed to produce a trajectory in hidden dynamic space, x_j. The target learning method can be applied to a wide variety of smoothing functions. The filter used for the smoothing in this paper is a second-order symmetrical (forward-backward) low-pass filter, whose single time-constant parameter, p_j, is also determined by the phone class. In the multidimensional case, there is a different time-constant for each dimension of the hidden dynamic space (the motivation for this will be described later).

The hidden dynamic trajectory is mapped to the surface acoustic form, y_j, by a multidimensional non-linear mapping, here a multilayer perceptron (MLP). A single MLP is used for all phones, and so this mapping defines the hidden dynamic space. This mapping can be considered analogous to the mapping between vocal tract shapes and speech sounds, although we intend it to be learned only from the acoustic data (as described shortly), and not restrict it to any such predetermined form.

Training of the HDM (described in [11]) can be carried out by comparing training data with the synthetic speech patterns produced by the HDM. All the HDM parameters are optimized by gradient descent to improve this fit.

Figure 2 shows hidden dynamic trajectories and synthetic acoustic patterns produced at various stages in such a gradient descent learning procedure. The training data is an 11 second connected vowel utterance which has been manually segmented and labelled with six phone classes (one referring to silence). The MLP used in the HDM here has 40 units in one hidden layer, and the time constant parameters are fixed during training.

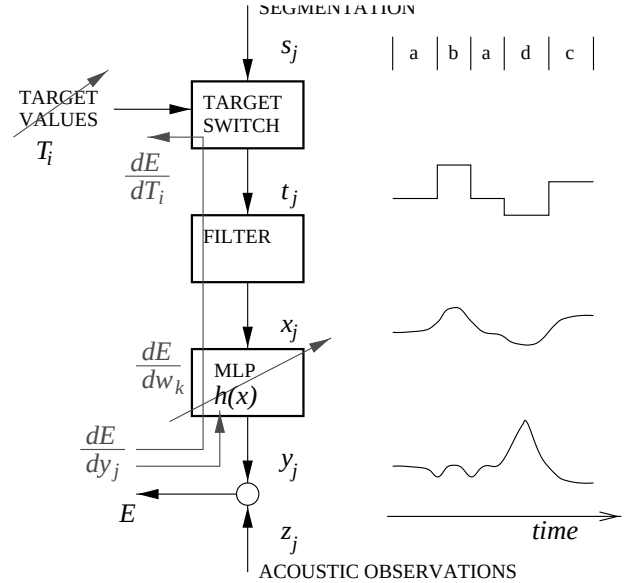

Figure 1: Calculating an acoustic error using the Hidden Dynamic Model. Derivatives of this error are backpropagated to the MLP weights and the target vectors.

5. THE HIDDEN SPACE

Because all the time constant parameters are equal in the HDM described in the previous section, the hidden dynamic space can be an arbitrary linear transform of these $\underline{x}$ parameters to give the same results. In Figure 3(a), the $\underline{x}_j$ trajectories have been offset and scaled to show that they closely resemble the formant frequencies (on a warped frequency scale).

The hidden dynamic trajectory is plotted in the original unscaled hidden space in Figure 3(b). The vertices of this trajectory correspond to the phone targets, and because the vowels have been produced quite slowly in this utterance, the trajectory approaches very close to the target for each vowel. It can be seen that the x-axis corresponds to F_1, while the y-axis corresponds to F_2.

Also of interest here is the way in which the hidden dynamic model encodes the silence 'phone' in the hidden representation, using a undefined region of formant space where F_2 is below F_1. Of course, in general it would be more appropriate to increase the dimensionality of the hidden space so that amplitude could become an independent parameter.

6. IMPORTANCE OF NON-LINEARITY

Using a linear mapping in place of the MLP in the HDM is equivalent to using the simple transitions given by the dynamic system, but operating directly in the acoustic domain. Simple, monotonic transitions in the spectral domain are insufficient to produce realistic speech patterns, even for vowel-vowel transitions [12].

This can be confirmed by considering a normal [ai] transition where the second formant sweeps up from about 1100Hz to 2300Hz. If a frequency band somewhere between these two frequencies is considered, then the amplitude will increase as the formant enters the frequency band, and decrease again as it leaves

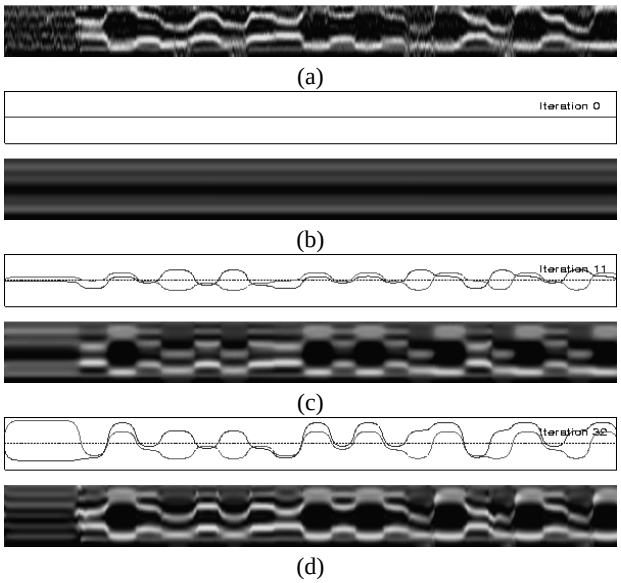

Figure 2: The training acoustic pattern $\underline{z}_j$ (a), and the hidden dynamic trajectory $\underline{x}_j$ and synthetic speech $\underline{y}_j$ at different points in the training procedure: (b) at initialisation, (c) after 11 iterations, and (d) after 32 iterations. The spectrograms shown have been derived from the 12th order MFCC representation used for $\underline{y}_j$ and $\underline{z}_j$.

it. This non-monotonic transition cannot be produced using just two target vectors (for the [a] and [i]), and target-directed linear dynamics.

Figure 4 shows synthetic speech produced from an HDM, which has been trained with the same training data as in Section 4, but uses a linear mapping in place of the MLP used in Section 4. As expected, the linear system is unable to reproduce convincing connected formant transitions like those seen in Figure 2(d), and simply cross-fades between spectrum shapes. This demonstrates the importance of the non-linear mapping in modelling the dynamics of the speech pattern.

7. DYNAMICS

Details of the time-varying symmetrical smoothing that constitutes the dynamic part of the HDM can be found in [11]. Figure 5 shows examples of trajectories which can be generated using such a versatile dynamic system. The symmetrical (non-causal) nature of the filter allows us to use intuitive segment boundaries, placed within the transition as opposed to preceding it.

The per-segment time-constants can also be thought of as target importance weights, variances associated with the targets, or the strength of influence of the targets on the trajectory. Figure 5 shows that by varying the time-constant for a segment, the trajectory can either approach the the target closely, or almost ignore it.

We chose this form of dynamics so that it would be useful for some aspects of consonant gestures. For example, in the production of a bilabial consonant such as the voiced stop, /b/, the most important articulatory action is to make a closure at the lips. The shape of the remainder of the vocal tract, such as the tongue position, is determined mainly by the context in this case [13]. According to the 'critical articulator' theory [14], the production of other consonants is similar, with the consonant influencing usually only a local region of the vocal tract. This is in contrast to the way vowel sounds specify an overall shape to the vocal tract.

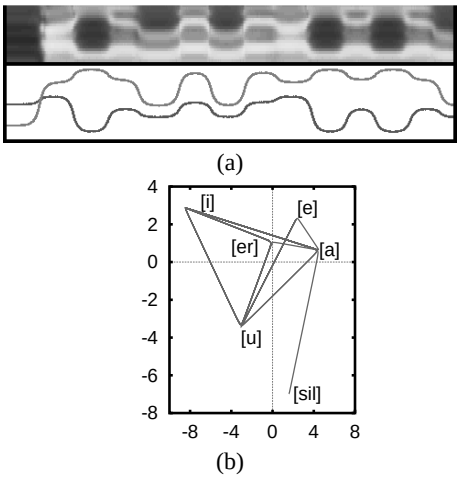

Figure 3: The synthetic spectrogram ((a) top) and the two hidden dynamic parameters ((a) bottom and (b)). In this case, the HDM has discovered the formant frequencies on a warped frequency scale.

Although the hidden representation of the HDM is an abstract hidden space, which is derived only from the data used to train the HDM, the time constant parameters, P_i, at least give it the flexibility to synthesize the sort of trajectories that we believe occur in articulatory space.

8. DIMENSIONALITY OF THE HIDDEN SPACE

The two-dimensional hidden space used for the experiment shown in Figure 2 is clearly insufficient for speech in general, although it is not clear what order is required.

We could postulate that one parameter is necessary for each place position, or each formant frequency, with perhaps extra parameters for amplitude, etc. A principled approach would be to train HDMs with different hidden-space dimensionalities, and select the one which generalizes the best. A similar method could be used to determine the necessary complexity of the MLP.

Figure 6 shows the approximation of the training data used in Figure 2 using a HDM with one, two and three parameters in the hidden representation. While the HDM is clearly unable to reproduce

Figure 4: Synthetic speech produced by an HDM with using a linear mapping instead of an MLP. The HDM here has been trained using the same material as in Figure 2.

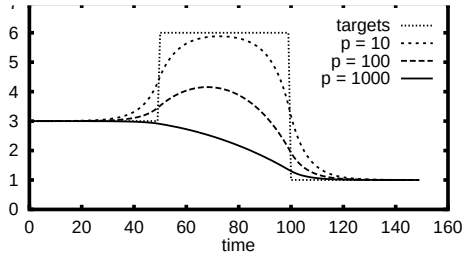

Figure 5: Hidden dynamic trajectories with time constants of 10, 100 and 1000 units for the centre segment.

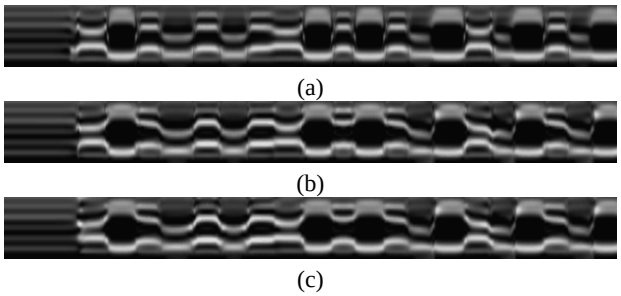

Figure 6: Synthetic speech produced by an HDM with using (a) one, (b) two, and (c) three dimensions in the hidden space. The HDM here trained using the same material as in Figure 2.

speech patterns with a single hidden dimension, there is little extra to be gained (for this simple vowel data) in using more that two parameters. Of course, for speech in general, we expect that a greater hidden space dimensionality will be required. In the work reported in [15] we typically used 6.

9. CONCLUSIONS AND FUTURE WORK

We think that the simple, flexible structure described here has potential for capturing important aspects of the relationship between phonetic labels and acoustic patterns, with potential applications in speech science, synthesis and recognition. Because it is so simple, there are also many possibilities for extending it.

As mentioned previously, it may be possible to enhance the model by allowing stochastic variation in not only the acoustic observations, but also the target values for a given segment. Ideally this variation could also be multimodal to accommodate distinctly different target values for the same 'phone'.

At present we rely on hand segmentations (or those provided by a conventional ASR system). We are keen to introduce a method of automatically adjusting segment boundaries. Because of the continuous nature of the hidden space, dynamic programming is not applicable to HDMs, but several possibilities for sub-optimal alignment search exist.

Because of the economical parameterisation, there are some interesting possibilities for dealing with speaker differences. For instance we can hope that most of the variation in the sets of target vectors for different speakers will be included in a low-dimensional 'speaker space', in which speaker adaptation can be attempted [7] [16].

We have some hope of extending the model to be compatible with modern phonological theories based on overlapping features, as proposed by Deng [17].

10. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. (#IIS-9732388), and was carried out at the 1998 Workshop on Language Engineering, Center for Language and Speech Processing, Johns Hopkins University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or JHU.

11. REFERENCES

[1] J. Clark and C. Yallop. *An introduction to phonetics and phonology*. Blackwell, Oxford and Cambridge, MA, 1990.

[2] M. J. Russell. A Segmental statistical model for speech pattern recognition. In *Proc. ICASSP'93*, 1993.

[3] M. J. F. Gales and S. J. Young. Segmental Hidden Markov Models. In *Proc. EuroSpeech-93*, pages 1579–1582, 1993.

[4] V. Digalakis, J. R. Rohlicek, and M. Ostendorf. A dynamical system approach to continuous speech recognition. In *Proc. ICASSP-91*, volume 1, pages 289–292, 1991.

[5] W. J. Holmes and M. J. Russell. Speech recognition using a linear dynamic segmental HMM. In *Proc. EuroSpeech-95*, pages 1611–1614, 1995.

[6] J. N. Holmes, I. G. Mattingly, and J. N. Shearme. Speech synthesis by rule. *Language and Speech*, 7:127–143, 1964.

[7] J.S. Bridle and M.R. Ralls. An approach to automatic speech recognition based on synthesis-by-rule. In F. Fallside and W. A. Woods, editors, *Computer speech processing.* Prentice Hall, 1985.

[8] L. Deng, P. Kenny, M. Lennig, V. Gupta, and P. Mermelstein. A Locus model of coarticulation in an HMM speech recognizer. In *Proc. ICASSP-91*, pages 97–100, 1991.

[9] R. Bakis. Coarticulation modeling with continuous-state HMMs. In *Proc. IEEE Workshop Automatic Speech Recognition*, pages 20–21, Arden House, New York, 1991.

[10] C. S. Blackburn and S. J. Young. Towards improved speech recognition using a speech production model. In *Proc. Eurospeech-95*, volume 3, pages 1623–1626, 1995.

[11] H. B. Richards and J. S. Bridle. The HDM: A segmental hidden dynamic model of coarticulation. In *Proc. ICASSP-99*, 1999.

[12] H. B. Richards. Articulatory parameters for speech analysis. *Ph.D. Thesis, University of Wales Swansea*, 1997.

[13] S. E. G. Öhman. Numerical model of coarticulation. *J. Acoust. Soc. Am.*, 41:310–320, 1967.

[14] 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.

[15] J. Bridle, L. Deng, J. Picone, H. Richards, T. Kamm, Z. Ma, M. Schuster, S. Pike, and R. Reagan. Final report on dynamic segmental models of speech coarticulation. In *NSF Workshop on Language Engineering (see also http://www.clsp.jhu.edu/ws98/)*, 1998.

[16] R. Kuhn, P. Nguyen, J. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini. Eigenvoices for speaker adaptation. In *Proc. ICSLP-98*, 1998.

[17] L. Deng. A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. *Speech Communication*, 24(4):299–323, 1998.