Kalman-filter based Join Cost for Unit-selection Speech Synthesis

Jithendra Vepa, Simon King

Centre for Speech Technology Research
University of Edinburgh
Edinburgh, UK
vepa@cstr.ed.ac.uk, Simon.King@ed.ac.uk

Abstract

We introduce a new method for computing join cost in unit-selection speech synthesis which uses a linear dynamical model (also known as a Kalman filter) to model line spectral frequency trajectories. The model uses an underlying subspace in which it makes smooth, continuous trajectories. This subspace can be seen as an analogy for underlying articulator movement. Once trained, the model can be used to measure how well concatenated speech segments join together. The objective join cost is based on the error between model predictions and actual observations. We report correlations between this measure and mean listener scores obtained from a perceptual listening experiment. Our experiments use a state-of-the art unit-selection text-to-speech system: *rVoice* from Rhetorical Systems Ltd.

1. Introduction

Unit-selection based speech synthesis systems attempt to select optimal speech units from large database, typically containing many instances of each speech unit with varied prosodic and spectral characteristics. Then, these speech segments are concatenated to produce high quality synthetic speech. The selection of the best unit sequence from the database is based on a combination of two costs: target cost (how closely candidate units in the inventory match the required targets) and join cost (how well neighbouring units can be joined)[1]. The optimal unit sequence is then found by a Viterbi search for the lowest cost path through the lattice of the target and join costs.

The ideal join cost is one that, although based solely on measurable properties of the candidate units, such as spectral parameters, amplitude and F0, correlates highly with human perception of discontinuity at unit concatenation points. A few recent studies [2, 3, 4, 5] have attempted to determine which objective distance measures are best able to predict audible discontinuities. Most of these studies focused on human detection of audible discontinuities in **isolated words** generated by concatenative synthesisers. We extended this work to the case of **polysyllabic words in natural sentences** and new spectral features, Multiple Centroid Analysis (MCA) coefficients [6]. But we found that no single distance measure performed well for all cases. A measure weighting individual MCA coefficients gave the best result [7], achieving more (seven out of ten cases) significant correlations with the perceptual data than any other measure. We believe there is still considerable room for improvement, and have developed a distance measure based on linear dynamical models.

In this paper, we propose the use of a **learned underlying representation** to define a join cost. The linear dynamical model is a probabilistic, continuous state-space model which infers a predicted (smooth) trajectory through a series of noisy observations by making smooth, continuous motion in a hidden state space, which is then projected up to the observation space. The model can simultaneously infer the most likely observation trajectories and compute the probability of the actual noisy observations. In this paper we use this probability as the basis of a join cost.

After units are concatenated, most systems attempt some form of local parameter smoothing to disguise the remaining discontinuity. The join cost measure and the join smoothing method interact closely. If we had a sufficiently large database and a perfect join cost measure then no smoothing would be required. Conversely, if we could smooth joins better, then the method of computing join cost would be less critical and a smaller inventory might be possible. We propose combining join cost computation and join smoothing – the work presented here is a first step towards this.

2. Linear Dynamical Model

A linear dynamical model (LDM) is described by:

$$\mathbf{y}_t = H\mathbf{x}_t + \epsilon_t \qquad \epsilon_t \sim N(\mathbf{v}, C) \qquad (1)$$

$$\mathbf{x}_t = F\mathbf{x}_{t-1} + \eta_t \qquad \eta_t \sim N(\mathbf{w}, D) \qquad (2)$$

where $\mathbf{y}_t$ is an observed feature vector, $\mathbf{x}_t$ is an unobserved (hidden) state vector with initial value at $t = 0$ of $\mathbf{x}_0$, ϵ_t and η_t are uncorrelated normally distributed noise vectors with means $\mathbf{v}, \mathbf{w}$ and covariance matrices C, D respectively. Recently, this model has been used for speech recognition [8], formant tracking [9] and estimation of vocal tract parameters [10].

LDMs learn an underlying, typically low dimensional, state space to model seemingly complex behaviour in observation space.

At present, our models are phone specific, with one set of parameters $H, F, C, D, \mathbf{v}, \mathbf{w}$ and $\mathbf{x}_o$ per phone. This is convenient, since joins are usually at the centres of phones in concatenative synthesis, so we can "run" a model from the start of a phone, through the join, to the end of that phone.

To understand how this helps us compute how close to natural the joined phone is, it is helpful to think of the LDM tracking the observation (LSF) trajectories. At the start of the phone, the model will follow these trajectories closely since the speech is natural – i.e. it is very much like the data the model was trained on; around the join, the model will deal with any discontinuity in the LSF trajectories as noise and infer a smooth path through the join – the error between this smooth path and the actual observations forms the basis of our join cost; towards the end of the phone, the model once again follows the observations closely since the speech is natural.

2.1. The EM Algorithm

The models are trained on natural speech: they will learn the dynamical properties of LSFs from natural examples of a particular phone. We use the Expectation-Maximisation (EM) algorithm to compute maximum-likelihood estimates for the model parameters $\{H, F, \mathbf{v}, \mathbf{w}, C, D, \mathbf{x}_0\}$. During the E step, statistics are accumulated over these training examples using the previous set of model parameters. Then in the M-step, these statistics are used to update the model parameters. Refer to [11] for full details.

The models typically needed 3-4 iterations for EM to converge. We experimented with three different schemes for initialising model parameters prior to EM. They are:

- **AR(1)**: A first order autoregressive (AR) process with some randomness introduced into the estimation, using a modified version of the software presented in [12].
- **Factor Analysis**: A factor analysis model is used to initialise the observation process parameters (H, v, C) which are then used to infer the state-space equation parameters (F, w, D).
- **Empirical**: Hand-picked initial values for the model parameters (refer to [13] for more details).

3. Perceptual Listening Tests

To evaluate our objective join costs, we use data from a perceptual experiment. A listening test was designed to measure the degree of **perceived** concatenation discontinuity in sentences synthesised by a state-of-the-art speech synthesis system (*rVoice*), using an adult North-American male voice. A preliminary assessment indicated that spectral discontinuities are particularly prominent for joins in the middle of diphthongs, presumably because this is a point of spectral change. Our study therefore focused on such joins. We selected two natural sentences for each of five American English diphthongs (ey, ow, ay, aw and oy), listed in our previous papers [6, 7].

3.1. Test Preparation

These sentences were then synthesised using the experimental version of *rVoice* speech synthesis system. For each sentence we made various synthetic versions, by varying the two diphone candidates which make the diphthong and keeping all the other units the same. We removed the synthetic versions which had poor joins in the neighbouring phones to the diphthong. The remaining versions were further pruned based on target features of the diphones making the diphthong, to ensure similar prosody among synthetic versions. This process resulted in around 30 versions with variation in concatenation discontinuities at the diphthong join. The authors manually selected what they judged to be the best and worst synthetic versions by listening to these 30 versions. This process was repeated for all ten sentences in our stimuli.

3.2. Test Procedure

There were around 17 participants in our perceptual listening test, most of them were PhD or MSc students with some experience of speech synthesis. Most of them were native speakers of British English.

Subjects were first shown the written sentence, with an indication of which word contains the join. At the start of the test they were first presented with a pair of reference stimuli: one

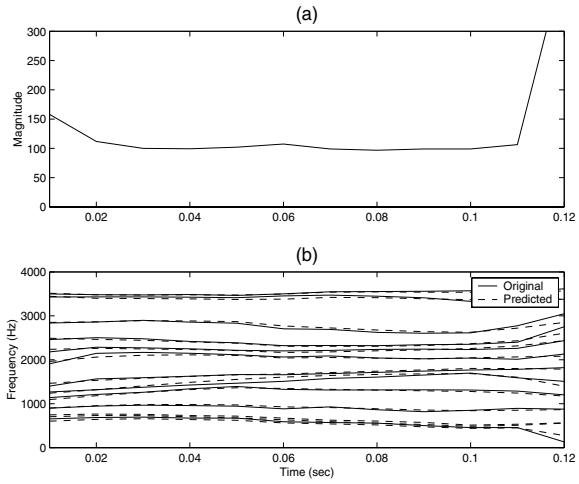

Figure 1: (a) Negative log likelihood estimate for a good join, (b) Corresponding original and predicted observations

containing the best and the other the worst joins (as selected by the authors) in order to set the end points of a 1-to-5 scale. Subjects could listen to the reference stimuli as many times as they liked and they could also review them at regular intervals (every 10 test stimuli) throughout the test.

They were then played each test stimulus in turn and were asked to rate the quality of that join on a scale of 1 (worst) to 5 (best). They could listen to each test stimulus up to three times. Each test stimulus consisted of first the entire sentence, then only the word containing the join (extracted from the full sentence, not synthesised as an isolated word).

The test was carried out in blocks of around 35 test stimuli, with one block for each sentence. Subjects could take as long as they pleased over each block, and take rests between blocks. Each test block contained a few duplications of some test stimuli to validate the subjects scores, as explained in section 5.

4. Objective measure

We compute the log likelihood of the observation sequence $\mathbf{Y}$, given the parameters of model m as follows [11]:

$$log\ p(\mathbf{Y}|m) = -\sum_{t=t_{start}}^{t=t_{end}} \{log|\Sigma_{\mathbf{e}_t}| + \mathbf{e}_t^T \Sigma_{\mathbf{e}_t}^{-1} \mathbf{e}_t\} + const. \quad (3)$$

$\mathbf{e}_t$ and $\Sigma_{\mathbf{e}_t}$ are the prediction error and its covariance for model m, and can be obtained from the standard Kalman filter recursions.

The upper halves of figures 1 and 2 plot the estimate of $-p(\mathbf{Y}|m)$ for a good join and poor join respectively. The increased model prediction error can be clearly seen in figure 2 where, in the region of the join, the model infers a smooth trajectory through the LSF parameters (lower half of the figure) but accumulates error between this smooth trajectory and the actual observations.

An objective join cost measure can be derived from the shape of the negative log likelihood plot. We have tried three different methods: 1) an average of the negative log likelihood estimates over 5 frames centred on the join; 2) the relative increase (over an estimated baseline) in the negative log likelihood estimates averaged over 3 frames centred on the frame

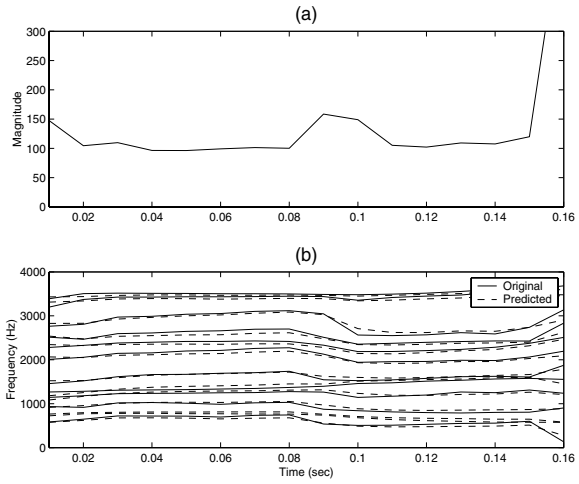

Figure 2: (a) Negative log likelihood estimate for a bad join, (b) Corresponding original and predicted observations

with highest estimate; 3) same as measure 2, except over 4 frames. Measures 2 and 3 are motivated by the plots in the upper halves of figures 1 and 2 – they measure the area of the "lobe" in figure 2.

5. Results and Discussion

In table 1, we present the number of subjects who listened to each sentence, and the number of subjects with more than 50% consistency in rating the joins. This consistency was measured on a validation set, which we included in the test stimuli for each sentence. We also manually checked all the listeners' ratings, and removed listeners who tended to always give the same rating (e.g all '1's), since this would not be caught by the consistency check. The mean listener scores were then computed only for the remaining subjects with more than 50% consistency in rating the joins.

	no. of subjects	consistent subjects
ey	13, 14	11, 8
ow	11, 13	6, 7
ay	17, 11	9, 6
aw	11, 13	11, 10
oy	13, 14	6, 6

Table 1: Consistency of subjects in listening tests, each number in a pair corresponds to each sentence.

Tables 2, 3 and 4 report the correlation coefficients of the three types of our analytical measures using likelihood estimates with mean listener preference ratings. The values 8k and 16k are waveform sampling frequencies. Correlation coefficients above the 1% significance level have been highlighted.

We found that training the model on only 10% of the 3400 available sentences is as good as training on the full data. Hence, we present those results only. Also, choosing the correct dimension for the state space is very important. We trained models with various state dimensions from 1 to 12. The correlations with perceptual ratings do not show any consistent trend: in some cases, a low state dimension yields high correlations; in other cases, a higher state dimension was required. Hence, only the best result for each case is quoted in the tables.

	measure 1		measure 2		measure3	
	8k	16k	8k	16k	8k	16k
ey	-0.40	0.06	-0.44	0.00	-0.42	-0.07
	-0.31	-0.18	-0.34	-0.20	-0.52	-0.50
ow	0.27	0.11	0.09	0.20	0.03	0.28
	0.43	0.17	0.04	0.19	0.06	0.11
ay	0.20	**0.51**	0.32	**0.52**	0.32	**0.48**
	0.32	**0.62**	-0.01	**0.45**	0.20	**0.46**
aw	-0.14	-0.13	-0.09	0.14	0.00	-0.12
	0.45	0.36	0.22	0.35	**0.49**	0.31
oy	-0.06	0.05	-0.19	-0.12	-0.19	0.05
	-0.08	-0.08	-0.09	-0.11	-0.02	-0.07

Table 2: Correlations between perceptual scores and three measures based on a LDM estimate, using the AR(1) method to initialise the model parameters prior to EM.

	measure 1		measure 2		measure3	
	8k	16k	8k	16k	8k	16k
ey	-0.28	**0.58**	-0.37	**0.56**	-0.32	0.33
	-0.27	-0.17	-0.28	-0.19	-0.51	-0.43
ow	0.25	0.26	0.11	**0.47**	0.10	**0.52**
	0.44	0.34	0.17	0.34	0.22	0.17
ay	0.34	**0.56**	**0.43**	**0.50**	**0.41**	**0.43**
	0.39	**0.59**	0.15	**0.44**	0.34	**0.49**
aw	0.07	-0.02	-0.02	0.13	0.08	-0.08
	0.55	**0.50**	0.28	0.43	**0.64**	0.42
oy	0.39	**0.45**	0.21	0.29	0.24	0.34
	-0.14	-0.14	-0.11	0.10	-0.06	0.03

Table 3: Correlations between perceptual scores and three measures based on a LDM estimate, using a factor analyser to initialise the model parameters prior to EM.

Comparison of tables 2, 3 and 4 shows that initialising the model parameters using a factor analyser or our hand-picked values yields better correlations than using an AR(1) model. Also, it is clear that measure 1 is best among our three analytical measures. This measure uses a simple average of the absolute model error over 5 frames centred on the join. The other measures attempt to calculate the "extra" error – the lobe seen in the upper half of figure 2 – which we hypothesised would be a good indicator of the difference between the joined phone and a natural token. This hypothesis is not supported by our results.

In table 5, correlations obtained using our previous measures [6, 7] based on statistical differences of spectral features are shown alongside those using the best of the new methods presented in this paper: measure 1 with models initialised using factor analysis. The second column (MFCC) shows correlations of perceptual scores with a Mahalanobis distance between Mel frequency cepstral coefficients, the third column (LSF) is for a Mahalanobis distance between line spectral frequencies and their deltas. The fourth column (MCA) shows correlations with absolute distances between multiple centroid analysis parameters and their deltas. The next column (MCA wgts.) presents the correlations with absolute distances between weighted MCA parameters. From these results we observe that our new method performs better than MFCC and LSF, and as good as MCA pa-

	measure 1		measure 2		measure3	
	8k	16k	8k	16k	8k	16k
ey	-0.26	**0.59**	-0.28	**0.59**	-0.27	0.39
	-0.27	-0.16	-0.27	-0.18	-0.50	-0.43
ow	0.27	0.27	0.11	**0.50**	0.008	**0.55**
	0.58	0.35	0.25	0.38	0.27	0.30
ay	0.33	**0.55**	**0.47**	**0.64**	**0.46**	**0.64**
	0.43	**0.55**	0.18	0.42	0.32	0.35
aw	0.08	-0.03	0.14	0.13	0.19	-0.07
	0.64	**0.50**	**0.59**	**0.46**	**0.73**	**0.46**
oy	0.41	**0.45**	0.22	0.37	0.27	0.43
	0.19	0.06	0.20	0.15	0.30	0.13

Table 4: Correlations between perceptual scores and three measures based on a LDM estimate, using hand-picked values to initialise model parameters prior to EM.

rameters. But, weighted MCA distances are best among all these measures. However, we believe there is still lot of scope for improvement using our linear dynamical model approach.

	MFCC	LSF	MCA	MCA wgts.	LDM
ey	0.21	0.37	0.36	**0.44**	**0.58**
	0.66	**0.58**	**0.46**	**0.60**	0.17
ow	0.31	0.21	0.19	0.19	0.26
	0.56	0.40	**0.46**	**0.52**	0.34
ay	0.39	0.01	0.03	-0.02	**0.56**
	0.66	**0.61**	**0.45**	**0.49**	**0.59**
aw	0.34	**0.66**	0.35	**0.49**	-0.02
	0.77	**0.78**	**0.57**	**0.62**	**0.50**
oy	0.17	0.28	**0.53**	**0.55**	**0.45**
	-0.01	0.17	0.30	0.39	-0.14

Table 5: Comparison with our previous results – see text.

6. Conclusions and future work

Further research is needed to get good analytical measures based on the LDM log likelihood estimate. Each diphthong, and indeed each of the two sentences for a single diphthong, required a different state-space dimension for peak performance. We chose the state-space dimension empirically using perceptual data, so to extend this technique from diphthongs to all segment types would require significant amounts of perceptual data. The fact that different segments show markedly different performance for any given join cost, and the fact that no single join cost performs well in all cases, indicate that finding a universal join cost is a hard problem.

6.1. Join smoothing

One of our motivations for using LDMs is that, as well as computing the join cost, they are able to smooth the LSF coefficients. This smoothing, we believe, should be better than *ad hoc* interpolation because a) the LSFs are treated as a set of parameters (an observation **vector**) and not independently, and b) the degree and extent of smoothing is controlled by the model parameters which are learned from natural speech – in other words, the model knows how natural LSFs behave, and attempts to make the joined LSFs look similar. To evaluate this, we are planning to do further listening tests using speech synthesised from smoothed LSFs inferred using the LDM, and compare this to simple interpolation.

7. Acknowledgements

Thanks to all the experimental subjects: the members of CSTR, staff at Rhetorical Systems Ltd. and students on the M.Sc. in Speech and Language processing, University of Edinburgh. The authors also thank Joe Frankel of CSTR for providing initial parameter values for EM, and also for his valuable suggestions throughout this study.

8. References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, 1996, pp. 373–376.

[2] E. Klabbers and R. Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," in *Proc. ICSLP*, 1998, pp. 1983–1986.

[3] J. Wouters and M. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," in *Proc. ICSLP*, 1998, pp. 2747–2750.

[4] Y. Stylianou and Ann K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. ICASSP*, 2001.

[5] Robert E. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesisers," in *The 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2001.

[6] J. Vepa, S. King, and P. Taylor, "Objective distance measures for spectral discontinuities in concatenative speech synthesis," in *Proc. ICSLP*, Denver, USA, September 2002.

[7] J. Vepa, S. King, and P. Taylor, "New objective distance measures for spectral discontinuities in concatenative speech synthesis," in *Proc. IEEE 2002 WorkShop on Speech Synthesis*, Santa Monica, USA, September 2002.

[8] J. Frankel and S. King, "ASR - articulatory speech recognition," in *Proc. Eurospeech*, Aalborg, Denmark, September 2001, pp. 599–602.

[9] G. Smith, J. de Frietas, T. Robinson, and M. Niranjan, "Speech modelling using subspace and EM techniques," *Advances in Neural Information Processing Systems*, vol. 12, pp. 796–802, 1999.

[10] J. McKenna and S. Isard, "Tailoring Kalman filtering towards speaker characterisation," in *Proc. Eurospeech*, Budapest, Hungary, September 1999, pp. 2793–2796.

[11] V. Digilakis, J. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its applications to speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 4, pp. 431–442, October 1993.

[12] Z. Ghahramani and G. Hinton, "Parameter estimation for linear dynamical systems," in *Tech. rep. CRG-TR-96-2*, Dept. of Computer Science, Univ. of Toronto, 1996, Software at www.gatsby.ucl.ac.uk/~ zoubin/software.html.

[13] Joe Frankel, *Linear dynamic models for automatic speech recognition*, Ph.D. thesis, University of Edinburgh, forthcoming.