SPEECH CODING AND SYNTHESIS USING PARAMETRIC CURVES

Luis Miguel Teixeira de Jesus and Gavin C. Cawley
School of Information Systems,
University of East Anglia,
Norwich, U.K.
E-mail: {lmj,gcc}@sys.uea.ac.uk

Abstract

Accurate modeling of co-articulation, the context-sensitive merging of the boundaries between allophones in continuous speech, is vital for natural sounding speech synthesis. This paper describes initial research investigating the use of Bézier Curves to form models of co-articulation in human speech. A 12th order, pitch synchronous line spectral pair (LSP) [1] analysis is performed on a corpus of 239 phonetically balanced sentences of English speech. The resulting data are divided to form an inventory of the diphones occurring in the speech database. The trajectory of each line spectral pair parameter through each diphone can then be represented by a single cubic Bézier curve segment, found using the Levenberg-Marquardt curve fitting method [2, 3]. Results are presented showing the accuracy of Bézier models of the coarticulation between different types of speech sounds.

1 Introduction

Speech is produced as the result of a coordinated sequence of movements by articulators, such as the lips, tongue and jaw. For a given language, there are a finite number of elementary speech sounds, known as allophones, produced by these articulatory gestures. However, due largely to the physical inertia of the articulators, the boundaries between allophones in human speech are not distinct, instead there is a gradual transition from one speech sound to the next. This effect, known as coarticulation, must be reproduced in synthetic speech in order to produce a natural voice quality. Speech synthesis by rule systems, such as the Joint Speech Research Unit synthesiser [4] and MITalk [5], attempt to model the effects of coarticulation by interpolating formant parameters with a piecewise linear template, using interpolation parameters tabulated for each parameter for each allophone. The Holmes-Mattingley-Shearme [6] algorithm provides an early example of this approach. A set of ad-hoc rules that modify the interpolation parameters, where necessary, according to phonetic context is often required to achieve acceptable speech quality. Compiling and fine-tuning the tables of interpolation parameters for each allophone involves a great deal of painstaking manual comparison of the spectra of natural and synthetic utterances. Revoicing a text-to-speech synthesis system is therefore a costly operation.

This research aims to investigate the use of parametric curve fitting techniques to form a data-driven model the effects of coarticulation, based on cubic Bézier segments, without the extensive manual effort required to implement conventional models used in text-to-speech systems. The trajectories of each line spectral pair parameter between a pair of allophones can be represented by cubic Bézier segments, forming a model of the effects of coarticulation within a particular diphone. The parameters controlling the shape of the Bézier segments are determined using a least squares curve fitting procedure over examples of each diphone extracted from a phonetically annotated corpus of human speech. A suitable strategy must then be developed to blend adjacent Bézier segments together, according to phonetic context, to model the effects of coarticulation that extend beyond diphone boundaries. This paper describes the curve fitting procedure used to encode diphones using cubic Bézier segments and presents results of experimental work to determine the error of the models for each category of diphone.

1.1 The Bézier Curve

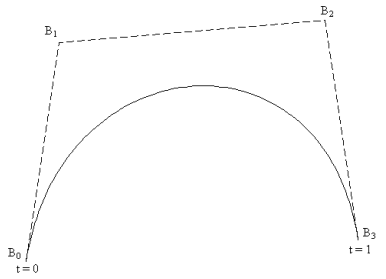

Figure 1: A cubic Bézier segment, also showing the defining polygon.

A Bézier segment is a parametric curve, defined by a polygon (a quadrilateral in the case of the cubic curves), as shown in figure 1. The polygon consists of two data points (B_0 and B_3) and one or more *control* points (B_1 and B_2); the curve passes through each of the data points, while the control points determine the initial direction of the curve at each data point. These curves were first defined, in terms of an initial point and a series of incremental vectors constituting successive sides of the defining polygon, by Pierre Bézier [7]. A. R. Forrest later developed the widely accepted formulation of the Bézier curve in terms of polygon vertices and recognized that the basis functions were Bernstein polynomials [8, 9]. Let $P(t)$ denote a Bézier curve of order n and B_i the i^{th} vertex of the defining polygon, then

$$P(t) = \sum_{i=0}^{n} B_i J_{n,i}(t) \qquad 0 \leq t \leq 1,$$

where

$$J_{n,i}(t) = \left(\begin{array}{c} n \\ i \end{array} \right) t^i (1-t)^{n-i}$$

In order to produce a smooth transition between Bézier segments, we need only ensure that the end point of each segment coincides with the start point of the next and that the neccessary first-derivative continuity conditions apply:

$$B_3 = B_{0(\text{next})},$$

$$B_{1(\text{next})} = 2B_{0(\text{next})} - B_0.$$

2 Method

The speech corpus used in this work consisted of 239 phonetically balanced sentences of neutrally articulated English speech, from a male speaker with a received pronunciation (RP) accent. The corpus provides raw speech data, in the form of 16bit linear samples at a sampling frequency of 16KHz and a time-aligned phonetic transcription based on the SAMPA phonetic alphabet [10] for each sentence. A twelfth order pitch-synchronous line spectral pair analysis of each sentence in the corpus was performed, using a default framelength of 30ms during unvoiced speech. The resulting frames of speech parameters were then divided to form an inventory of each of the diphones occurring in the corpus. To simplify the curve-fitting procedure, the duration of each example of each diphone was first normalised and then the frames of line spectral pair parameters resampled using simple linear interpolation, so each diphone is represented by ten frames of speech parameters. The Levenberg-Marquardt algorithm was then used to find a single cubic Bézier segment forming the best fit for each line spectral pair parameter, from the available examples of each diphone, and the root-mean-square (RMS) error recorded.

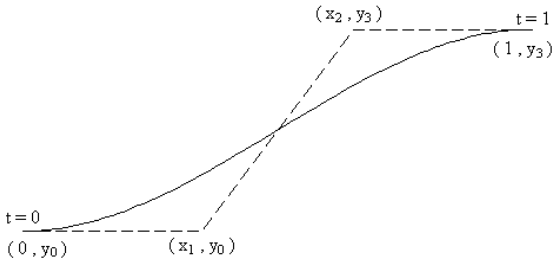

Figure 2: Constrained cubic Bézier segment used to model transitions in line spectral pair parameters between adjacent allophones.

In this study, in order to reduce the number of parameters controlling the model of each diphone, the constraint that the initial and final gradient of the Bézier segment is zero was imposed, as shown in figure 2 such that

$$y_0 = y_1 \qquad \text{and} \qquad y_2 = y_3$$

This constraint seems reasonable if the diphones are regarded as joining during steady state conditions at the centre of each allophone. This means that each diphone is represented by only four parameters y_0, x_1, x_2 and y_3 for each LSP parameter providing some amount of data compression.

3 Results

Table 1: Table of the numeric code assigned to each phonetic category of the set of allophones represented by the SAMPA alphabet.

Code	Category	Allophones
0	Silence	#:
1	Approximants	=l, l, r, w, j
2	Nasals	m, =m, n, =n, N
3	Plosives	b, p, d, t, g, k
4	Affricates and Fricatives	tS, dZ, v, f, D, T, S, s, z, Z, h
5	Vowels	I, E, {, V, Q, U, @, i, 3, u, A, O, I@, E@, U@, eI, aI, OI, @U, aU

Figure 4 shows the average root-mean-square error for each of 25 diphone categories. Each digit of the numeric code representing each category corresponds to the broad phonetic type of one of the allophones

forming the diphone, as shown in table 1, for example the diphone t E@ belongs to category 35 as the initial allophone is a plosive (group 3) and the final allophone is a vowel (group 5).

3.1 Blending Adjacent Diphones

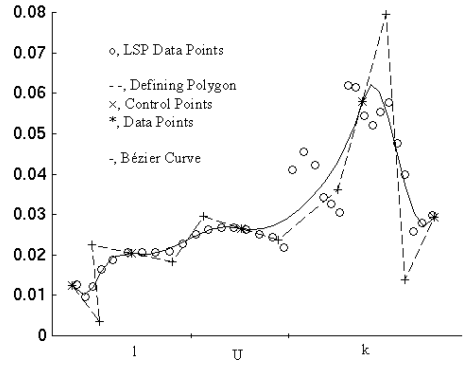

Figure 3: Trajectory of the first line spectral pair parameter for the word "look" modeled by Bézier segments, also showing the defining polygons.

Since diphones abut during the steady state conditions in the middle of each allophone, it seems sensible to blend the Bézier models of each diphone so as to ensure a smooth transition between each diphone, as described in section 1.1. Figure 3 shows the trajectory of the first line spectral parameter for the word "look" ("lUk" according to the SAMPA phonetic alphabet) generated from Bézier curve segments using this approach. The defining polygons are also shown. It can easily be seen that a single Bézier segment is able to model the transition between sonorants adequately; however, where abrupt transitions are required, principally during plosive sounds, a more complex model is needed. Plosive sounds are modelled in many synthesis-by-rule systems as a sequence of smaller speech sounds representing the closure, release and post-release phases so that the existing interpolation method can accommodate a more complex trajectory. This would also seem to be a sensible approach to adopt in this work.

4 Conclusions

Our initial research indicates that a single cubic Bézier curve can be used as an effective model of the trajectory for line spectral pair parameters during the transitions between a majority of speech sounds. Further research is needed to determine suitable strategies for blending Bézier segments to produce a natural sounding utterance.

5 Acknowledgments

The authors would like to thank Dr. Andy Breen at the British Telecommunications Laboratories at Martlesham Heath, Ipswich, U.K., for providing the speech data used in this research.

References

[1] N. Sugamura and F. Itakura. Speech analysis and synthesis methods developed at ECL in NTT — from LPC to LSP. In *Speech Communication*, volume 5, pages 199–215, 1986.

[2] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly Of Applied Mathematics*, II(2):164—168, July 1944.

[3] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal Soc. Indust. Applied Mathematics*, 11(2):431—441, June 1963.

[4] E. Lewis. *A 'C' implementation of the JSRU text-to-speech system.* Computer Science Department, University of Bristol, August 1989.

[5] J. Allen, M. S. Hunnicutt, and D. Klatt. *From text to speech: the MITalk system.* Cambridge University Press, 1987.

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

[7] P. E. Bézier. How Renault uses numerical control for car body design and tooling. In *Society of Automotive Engineer's Congress, SAE paper 680010*, Detroit, MI, 1968.

[8] A. R. Forrest. Interactive interpolation and approximation by Bézier polynomials. *CAD Computer-Aided Design, Special Issue: Bézier Techniques*, 22(9):527—537, November 1990.

[9] D. F. Rogers and J. A. Adams. *Mathematical Elements For Computer Graphics (2nd Edition).* MacGraw-Hill, 1990.

[10] SAMPA computer readable phonetic alphabet. http://www.phon.ucl.ac.uk/home/sampa/home.htm.

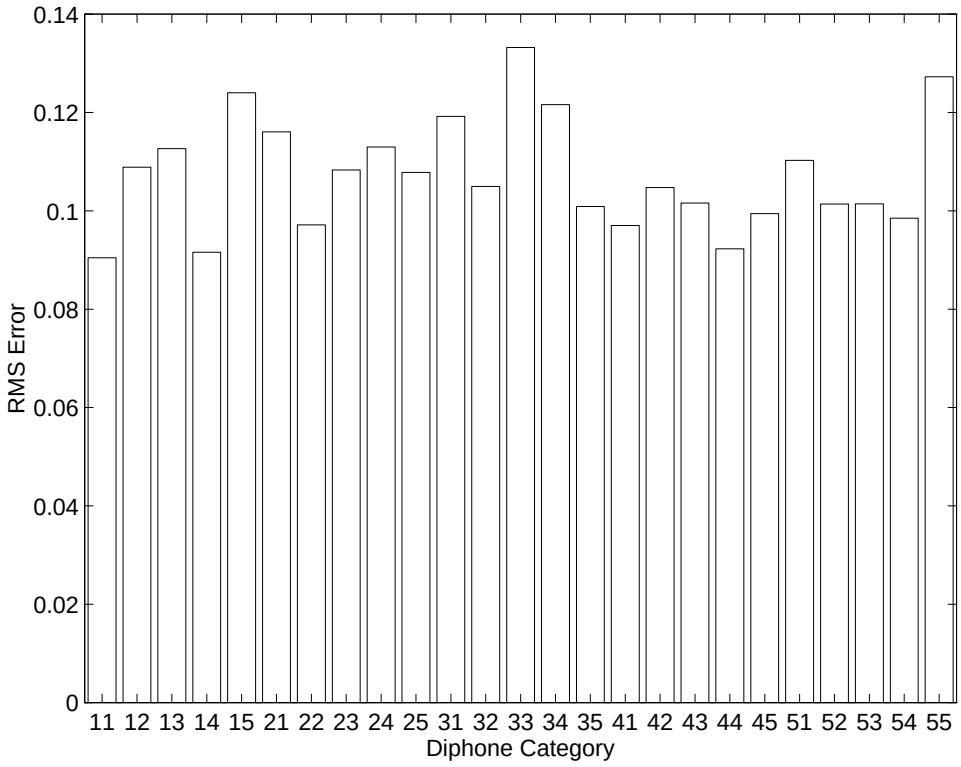

Figure 4: Bar chart showing root-mean-square error for Bézier models for each category of diphone