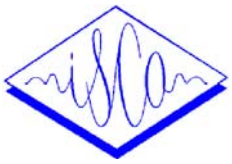


EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

MODELING SEGMENTAL DURATION WITH MULTIVARIATE ADAPTIVE REGRESSION SPLINES

Marcel Riedi

Institut TIK, ETH Zürich
8092 Zürich, Switzerland
riedi@tik.ee.ethz.ch

ABSTRACT

The application of "Multivariate Adaptive Regression Splines"(MARS) to the problem of modeling duration of a set of segments used in a text-to-speech system for German is presented. MARS is a technique to estimate general functions of high-dimensional arguments given sparse data. It automatically selects the parameters and the structure of the model based on data available. The result is a model with a correlation coefficient between observed and predicted durations of a test set of 0.90. Besides highly accurate predicting durations, a MARS model also allows interpretation of its structure, demonstrated in this study by analyses of factor importance and interactions of the MARS model.

1. INTRODUCTION

The goal of modeling segmental duration is to find a computational relation between a set of factors known to affect duration and the segment duration. The practical reasons are prediction and interpretation. In speech synthesis the primary interest lies in the models capability to predict natural sounding durations for all linguistically possible combinations of factor values. The interpretation of the model structure allows new knowledge about factor interactions to be gained.

"Multivariate Adaptive Regression Splines"(MARS) [1, 2] is a method to estimate general functions of high-dimensional arguments given sparse data. Its application to the problem of modeling segmental duration is described here. MARS uses data not only to calculate the parameter values of some function specified in advance, also the structure of the model is automatically determined. This structure consists of a selection of factors (among a set of factors initially specified) and their interactions.

Two often used methods of this automatic parameter and structure selection type are "Classification and Regression Trees"(CART) [3, 4] and "Neural Networks"mostly in the form of "Multilayer Perceptrons"(MLP) [5, 6, 7]. Objections raised against these two approaches are mainly insufficient accuracy of duration prediction in the case of CART, and the difficulties in interpreting the hidden structure of the model learned in the case of MLP.

As results show, with MARS durations can be highly accurate predicted. The MARS model permits human interpretation of the model structure, which suggests its use as data analysis tool. Other advantages of MARS are its ability to directly include categorical and ordinal factors in the model, automatic selection of the statistically most important factors and interactions, and inclusion of knowledge into the model building procedure.

2. MULTIVARIATE ADAPTIVE REGRESSION SPLINES

A detailed description of the theory and implementation of MARS is given in [1, 2]. A very brief overview of the ideas relevant for modeling segmental duration will be given here.

MARS estimates a model (spline function) of the form

$$\hat{f}(\mathbf{x}) = \sum_{m=0}^{M} \alpha_m B_m(\mathbf{x}), \qquad (1)$$

where $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ is a set of factor values and α_m is a real number indicating the contribution of the basis function

$$B_m(\mathbf{x}) = \prod_{k=1}^{K_m} b_{km}(x_{v(k,m)}).$$

$v(k,m)$ is the index of the factor used as argument of b_{km}. MARS is able to handle factors of ordinal (e. g. x_v some real number) and categorical (e. g. $x_v \in \{onset, nucleus, coda\}$) type in the same model. The b_{km} are defined in pairs:

$$\begin{aligned} b^{+}_{km}(x_{v(k,m)}) &= [x_{v(k,m)} - t_{km}]_+ \\ b^{-}_{km}(x_{v(k,m)}) &= [-(x_{v(k,m)} - t_{km})]_+ \end{aligned}$$

for ordinal factors , where t_{km} is a value in the same domain as $x_{v(k,m)}$. The function $[\cdot]_+$ equals its argument for positive arguments and zero otherwise. For categorical factors

$$\begin{aligned} b^{+}_{km}(x_{v(k,m)}) &= I(x_{v(k,m)} \in A_{km}) \\ b^{-}_{km}(x_{v(k,m)}) &= I(x_{v(k,m)} \notin A_{km}), \end{aligned}$$

where $I(\cdot)$ is an indicator function having value one if its argument is true and zero otherwise. A_{km} is a subset of the possible values of $x_{v(k,m)}$.

Starting with $B_0(\mathbf{x}) = 1$ the MARS learning algorithm iteratively adds pairs of basis functions to the model. These added basis functions consist of an earlier added basis function multiplied with a b_{km}^+ and b_{km}^- chosen to minimize the mean squared error (MSE) of the extended models output. An initially specified number of basis functions M_{max} are added, intentionally attempting to overly fit the data. Some of these basis functions are then deleted, the criterion being minimization of a modified generalized cross-validation (GCV) estimate of the MSE of the models prediction (GCV and other estimates are described in [1, 2]). Including available knowledge of the structural relations among the factors into the learning algorithm can be achieved by restricting the possible basis functions being added to the model. E. g., the maximum number K_{max} of factors involved in basis functions can be limited, or specific factor interactions can be disallowed.

By rearranging terms in Eq. (1), basis functions involving only one factor ($K_m = 1$) can be grouped together. The same can be done for basis functions involving two or more factors ($K_m \geq 2$). This rearrangement, named ANOVA decomposition, facilitates interpretation of the model. E. g., factors appearing only in the terms involving one factor contribute purely additive to the model, or two-factor interactions can be analyzed by displaying the contribution of each factor value combination to the model.

3. MODELING SEGMENTAL DURATION

Modeling segmental duration requires the choice of a set of factors to be included in the model. Initially, a redundant set of factors can be specified, and the MARS algorithm then automatically selects the ones contributing best to the model. The set of factors used in this study are (abbreviations given in parentheses for later reference):

- type of the segment described by a combination of length and position of tongue ($a0$), and the position of the first formant ($b0$) for vowels, respectively by a combination of articulation type and voiced/unvoiced distinction ($a0$), and the place of articulation ($b0$) for consonants,
- types of the two directly preceding ($a2\_$, $b2\_$, $a1\_$, $b1\_$) and following segments ($a1$, $b1$, $a2$, $b2$),
- position (absolute and relative to the nucleus) and number of segments in the syllable (sp, sc, sn),
- accentuation degree of the syllable (ac),
- position and number of syllables in the word (wp, wn) and foot (fp, fn),
- and position of the foot in the sentence (po).

Because of applying the MSE as criterion for minimization in the MARS learning algorithm, a normal probability distribution with zero mean of its prediction error is assumed. If this distribution considerably differs from a normal distribution a transformation of duration (e. g., logarithm) should be applied.

With a set of factors typically needed for modeling segmental duration one of the main problems is the sparsity of the data available for training (see [8], and for the data used in this study [7]). As demonstrated in [1], MARS is well suited for this sort of problems. A description of the natural speech data (21510 segments) used in this study can be found in [7].

Another problem encountered when modeling duration with, e. g., MLP is the requirement of ordered factors. For some factors no order can be found. In these cases some order has to be assumed, possibly resulting in inaccurate predictions. Solutions are the use of dummy variables for categorical factors or the separation of the segments to be modeled into subclasses (e. g., vowels and consonants) which can be modeled by ordinal factors (see [8]). The disadvantage of using dummy variables is the loss of the possibility to interpolate. The division into subclasses reduces the complexity of the models, but has the drawback of reducing the size of the training sample available for each of the subclasses. This possibly prevents successful modeling of structures valid for several subclasses.

MARS can directly handle both ordinal and categorical factors in the same model. Interpolation (smoothing) of categorical factors is done by building subsets of factor values with similar or equal effects. Building these subsets is done automatically, based on the data, thereby increasing the prediction accuracy of the model. Ordering factors and manually dividing the segments into subclasses provide ways to control the model building by utilizing outside-of-the-data knowledge. Methods for ordering factors are described in [8], using MARS for this purpose is another possibility.

The mean squared error used in the learning algorithm depends on the scale the prediction is measured in. Comparing the prediction accuracy of different models requires a scale independent measure. In this study an estimate of the relative mean squared error ($RMSE$)

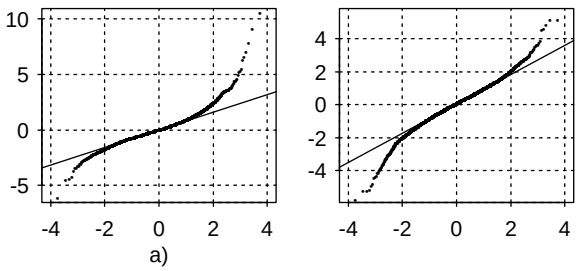
$$RMSE(\hat{f}) = \frac{MSE(\hat{f})}{MSE(\mu)}$$

was used. The $RMSE$ is the relative performance of the model compared to the simple model $\hat{f} = \mu$, where μ is the mean of the durations of the data available.

4. RESULTS

In all experiments 75% of the data available were used for calculating the MARS model and the other 25% made up the test set reserved for investigating the prediction accuracy (prediction errors, $RMSE$, and the correlation coefficient r). All of the above mentioned factors were used. No ordering of factors was attempted.

Normal probability plots of the prediction error (divided by its standard deviation) of MARS models are shown in Figure 1. As can be seen in plot a), directly modeling duration resulted in a skewed distribution of the prediction error, indicating the need for a transformation of duration. Modeling the logarithm of duration resulted in the desired symmetric distribution (cf. plot b)). A similar distribution resulted from modeling the duration to the power of $1/4$. The distribution is still leptokurtic which in general can not be corrected by a transformation of duration.



$M_{max} = 300$, $K_{max} = 5$). Duration has been modeled a) directly and b) after taking the logarithm.

Models with different maximum interaction levels K_{max} have been built to estimate the contribution of the different levels of interaction to the prediction accuracy. Table 1 displays the $RMSE$ calculated for models of the logarithm of duration only differing by K_{max}. Not allowing any factor interactions ($K_{max} = 1$) resulted in a $RMSE$ of 0.25. Allowing two-factor interactions considerably decreased the $RMSE$ to 0.20. By allowing three-factor interactions only a slight but still noticeable decrease of the $RMSE$ could be observed. Allowing even more than three-factor interactions increases the $RMSE$. On the other hand when directly modeling duration a minimal $RMSE$ of 0.21 was found for $K_{max} = 4$.

K_{max}	$RMSE$	K_{max}	$RMSE$
1	0.2545	4	0.2025
2	0.2048	5	0.2041
3	0.1989	19	0.2098

Table 1: The $RMSE$ for MARS models ($M_{max} = 300$) for different K_{max}.

Table 2 displays part of the ANOVA decomposition of a MARS model with $M_{max} = 600$, $K_{max} = 2$, and direct modeling of duration. 168 basis functions and a GCV of 306.8 remained after deletion of the ones not contributing to the prediction accuracy of the model. $GCV\_$ in Table 2 shows the GCV value of the model with the according group of basis functions removed. The basis functions with equal K_m and factor(s) are grouped together, with #bf being the number of basis functions in each group. All basis function groups involving a single factor and the ten two-factor groups with the highest contribution to the performance of the model are displayed. The factors $a0$, sn, fn, and ac appear as non-interacting factors, with $a0$ being the only one highly contributing to the models prediction accuracy. They are not purely additive since they also appear in terms involving two factors (sn and fn in the part of the ANOVA decomposition not shown here). Among the two-factor interactions $a0/a1$ had the highest contribution to the model.

$GCV\_$	#bf	factor(s)	$GCV\_$	#bf	factors
342.7	2	$a0$	317.1	7	$a0/b1$
309.8	1	sn	315.1	3	$b1/sp$
307.4	1	fn	312.7	4	$a0/a1\_$
307.0	1	ac	311.9	3	$b0/ac$
324.6	9	$a0/a1$	311.6	4	$a0/po$
322.7	6	$b0/b1$	311.4	4	$b0/a1$
318.5	5	$a0/b0$	311.0	2	$a1\_/a2\_$

Table 2: Part of the ANOVA decomposition of a MARS model with $M_{max} = 600$ and $K_{max} = 2$.

In the following the basis function groups $a0/a1$ and $a0/po$ will be discussed. The possible values of the factors $a0$ and $a1$ are:

lf:	long vowel, front	vf:	voiced fricative
lb:	long vowel, back	na:	nasal
sf:	short vowel, front	la:	lateral
sb:	short vowel, back	vi:	vibrant
up:	unvoiced plosive	dt:	diphthong
vp:	voiced plosive	pa:	preplosive pause &
uf:	unvoiced fricative		glottal closure.

Table 3 displays mapped durations for all factor value combinations of $a0$ and $a1$. These mapped durations are calculated by only using the $a0$ and $a0/a1$ basis function groups for prediction and mapping the resulting durations into the interval $[0,99]$. On the horizontal axis the factor values of $a0$ are displayed and the vertical axis shows the values of $a1$. It can be seen, e. g., that the value of an 'sf' segment followed by an 'up' segment is half the value of an 'lf' segment followed by an 'up' segment. Within the pairs (lf,lb), (sf,vi), and (uf,na) of $a0$ and the pair (lf,lb) of $a1$ the same effects can be observed. Referring only to the $a0$ and $a0/a1$ basis function groups these pairs could be combined to single values. With the exception of plosives, segments tend to be of shorter duration if followed by a nasal or lateral segment.

For a factor to be ordered its effects must always have the same direction. This does not hold for $a0$ and $a1$ as can be seen in Table 3: A nasal preceding a voiced plosive is longer than a lateral preceding a voiced plosive, whereas the opposite is true for the same type of segments followed by a voiced fricative. A short vowel with tongue position in the front is longer if followed by an unvoiced plosive than by a voiced fricative, but for a short vowel with tongue position in the back the opposite can be observed.

$a1^{a0}$	lf	lb	sf	sb	up	vp	uf	vf	na	la	vi	dt	pa
lf	90	90	47	56	11	7	56	44	56	53	47	93	53
lb	90	90	47	56	11	7	56	44	56	53	47	93	53
sf	84	84	44	51	7	4	51	41	51	48	44	88	48
sb	82	82	44	49	7	4	49	41	49	45	44	85	45
up	94	94	47	55	8	8	53	49	53	55	47	94	53
vp	93	93	47	56	7	3	56	44	56	53	47	97	53
uf	93	93	44	58	9	9	56	45	56	58	44	93	56
vf	94	94	42	60	8	8	58	43	58	60	42	94	58
na	72	72	34	50	12	12	49	36	49	50	34	72	49
la	71	71	31	51	7	4	51	27	51	48	31	75	48
vi	80	80	34	56	7	3	56	31	56	53	34	83	53
dt	84	84	47	51	11	7	51	44	51	48	47	88	48
pa	96	96	44	58	3	0	58	41	58	55	44	99	55

Table 3: Predicted durations (mapped into $[0, 99]$) for the factor value combinations of $a0$ and $a1$.

Mapped durations for all factor value combinations of $a0$ and po are displayed in Table 4. The values of po are:

si:	sentence initial	pf:	phrase final
sf:	sentence final	me:	neither initial nor final
pi:	phrase initial	mo:	phrase with one foot.

For the interaction of $a0$ and po larger clusters of $a0$ values (and fewer basis functions) can be observed than for $a0$ and $a1$: (lf,lb,dt), (sf,vf,vi), (sb,uf,na,la,pa), and (up,vp) for $a0$, and (si,pi) for po. As expected longer durations were observed for sentence and phrase final positions of the segment, as well as for segments in a phrase with a single foot.

po^{a0}	lf	lb	sf	sb	up	vp	uf	vf	na	la	vi	dt	pa
si	85	85	43	45	3	3	45	43	45	45	43	85	45
sf	91	91	43	51	3	3	51	43	51	51	43	91	51
pi	85	85	43	45	3	3	45	43	45	45	43	85	45
pf	99	99	48	51	0	0	51	48	51	51	48	99	51
me	89	89	45	45	0	0	45	45	45	45	45	89	45
mo	99	99	51	51	3	3	51	51	51	51	51	99	51

Table 4: Predicted durations (mapped into $[0, 99]$) for the factor value combinations of $a0$ and po.

By removing all basis function groups involving a certain factor from the model and calculating the GCV, the importance of this factor can be estimated. $a0$ has shown to be the most important factor followed by $b0$, $a1\_$, and $a1$. In Table 5 $a0$ is given the value 100, the other values are the relative importance of the factors compared to $a0$.

Models with different M_{max}, K_{max}, and transformations have been trained. The highest prediction accuracy ($RMSE = 0.1891$, correlation coefficient $r = 0.9008$) of a MARS model was achieved with a model of the duration to the power of $1/4$, $M_{max} = 300$, and $K_{max} = 3$. Informal acoustical tests with this model further demonstrated the accuracy of the predicted durations.

$a0$	100	$b1$	18	po	11	$b2\_$	6
$b0$	60	ac	18	fn	10	fp	6
$a1\_$	28	sp	14	sc	9	$b2$	5
$a1$	25	wp	14	$a2\_$	8	wn	2
sn	20	$b1\_$	12	$a2$	8		

Table 5: Relative factor importance ($M_{max} = 300$, $K_{max} = 3$, logarithm of duration modeled).

5. CONCLUSIONS

The MARS technique is well suited for modeling segmental duration for prediction and interpretation. It is language independent and should be equally well applicable for modeling syllable duration. The MARS learning algorithm is very computation intensive (from a few hours to several weeks on a Sparc 20 for this study). On the other hand, predicting duration with a MARS model has relatively low computational requirements, making it feasible to use such a model in a real-time speech synthesis system. Its performance in such an application will have to be compared to other models by formal acoustical tests. Possible improvements of the models described here include the use of other MSE estimates than the GCV, training of larger models, and ordering of the factors.

REFERENCES

[1] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–141, 1991.

[2] J. H. Friedman, "Estimating functions of mixed ordinal and categorical variables using adaptive splines," in *New Directions in Statistical Data Analysis and Robustness* (S. Morgenthaler, E. Ronchetti, and W. A. Stahel, eds.), pp. 73–113, Birkhäuser, 1993.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Chapman & Hall, 1984.

[4] M. D. Riley, "Tree-based modelling of segmental durations," in *Talking Machines: Theories, Models and Designs* (G. Bailly, C. Benoît, and T. R. Sawallis, eds.), pp. 265–273, North-Holland, 1992.

[5] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing* (D. E. Rumelhart and J. L. McClelland, eds.), vol. 1, ch. 8, pp. 318–362, MIT Press, Cambridge, Mass., 1986.

[6] W. N. Campbell, "Analog I/O nets for syllable timing," *Speech Communication*, vol. 9, pp. 57–61, 1990.

[7] M. Riedi, "A neural-network-based model of segmental duration for speech synthesis," in *Proc. Eurospeech'95*, vol. 1, pp. 599–602, ESCA, 1995.

[8] J. P. H. van Santen, "Assignment of segmental duration in text-to-speech synthesis," *Computer Speech and Language*, vol. 8, pp. 95–128, 1994.