TR-I-0276

Tree-based Unit Selection for English Speech Synthesis

Wern Jun Wang, W. N. Campbell,
Naoto Iwahashi & Yoshinori Sagisaka

ATR Interpreting Telephony Research Labs.

1992.8.4

Abstract

Data scarcity for many contexts is a serious problem in English speech synthesis. In this paper, we proposed a new unit selection scheme using broad categories for segment description and use regression trees to factorize the data. The cross-validation method was used to evaluate the validity of this algorithm. Experiment results show that regression trees offer a promising solution for data scarcity problem in the unit selection procedures for concatenative English speech synthesis.

Contents

1 Introduction

English speech synthesis system is now being developed at ATR. We are trying to incorporate Japanese non-uniform unit technique in such system. In the non-uniform unit synthesis system[1], the unit selection part is implemented by using four kinds of spectrum measures. These four measures are (1) contextual spectral difference (CSD) between source context and target context, (2) prototypicality of the segment for its context, (3) spectral gap between neighboring phoneme and (4) spectral discontinuity between adjacent segments. The feasibility of using these spectrum measures in English speech synthesis system has been investigated. According to the analysis results, we must find an alternative for CSD measure.

The reason is there exists some differences between Japanese and English. The main differences are the phoneme category number and consonant clustering problem. The CSD measure and CSD estimation that being employed in NUU system is not appropriate for English system. Although CSD estimation is designed for unknown context and it really works well in Japanese system, but it cannot work in sparse data domains because it relies on statistics of adjacent phone context effects derived from large numbers of samples. For English, the variety of context is often too large to obtain from the unit database. Therefore it is impossible to find all necessary spectrum for CSD measure and estimation.

Since every language has its own characteristics, to find language independent selection algorithm is a convenient way to overcome this data scarcity problem. The regression tree method has been adopted in this paper. The idea is to partition the spectrum space in terms of several broad classes to some homogeneous subsets. Then we could find some appropriate units by using the information at these homogeneous nodes.

In section 2, we will briefly describe the characteristics of regression tree analysis and the reason why it appears a promising candidate for this problem. The growing process of regression tree will be discussed in section 3 with evaluation by cross-validation and separability measurement in section 4. The last section will introduce the procedures for unit selection.

2 Tree Regression Analysis

Tree structured regression offers an interesting alternative for looking at regression type problems. The accuracy of regression trees has been competitive with linear regression. It can be much more accurate on non-linear problems but tends to be somewhat less accurate on problems with good linear structure[3]. Because tree-based models are more adapt at capturing nonadditive behavior; the standard linear model does not allow interactions between variables unless they are prespecified and of a particular multiplicative form[4].

This models are fitted by binary recursive partitioning whereby a dataset is successively split into increasingly homogeneous subsets until it is infeasible to continue.

The other advantages of regression tree algorithm are as follows: (1) statistically selects the most significant features involved, (2) provides "honest" estimates of its performance, (3) permits both categorical and continuous features to be considered, (4) allow human interpretation and exploration of the results[2].

Table 1: The common factors used in the speech analysis

features	cepstrum coefficients, log-power
analysis method	Improved cepstrum method
dimension	11
frame rate	5 ms
sampling frequency	12 kHz
environment factors	7 for vowel, 8 for consonant
distance measure	Euclidian distance
synthesis method	Log Magnitude Approximation Filter

Figure 1 shows a simple tree for predicting spectrum of the phone /a/. To know what the tree looks like first will be helpful to understand the following tree-growing process.

3 Regression Tree Generation

3.1 Description of Database

Our database includes 200 sentences utterred by a Britian female speaker. There are 8802 segments totally, includes 3223 vowels, 5414 consonants and 165 elision phonemes. The regression tree analysis will apply to 44 phonemes. In the evaluation, we will investigate the results of five basic vowels and some consonants, Table 1 shows the common factors used for the acoustic analysis of the speech data.

3.2 Three questions

Predicting with an existing tree is easy, the interesting question is how to generate the tree from a given set of input features. There are three basic questions that have to be answered when generating a regression tree.

1. What are the splitting criteria for each node ?

2. What are the stopping rule for each node and whole tree ?

3. How to predict the value at each node ?

Splitting criteria:

The coding of the phonetic context requires special considerations since a reasonable phone set contains many phones; In this case, we used 44 phones. If each segment position were treated as a single feature, about 2^{44} binary partitions would have to be considered for this variable at each node, clearly making this approach impractical. The solution adopted here is to classify each phone in terms of several broad classes[2], because we believe that a phoneme in an utterance is heavily dependent on its phonetic context, and the primary aspects of this coarticulation are the phonetic features of neighboring phonemes. For the regression tree of vowel, we use stress of current vowel and v/uv

Table 2: The descriptions of broad classes used in analysis

For vowel:	
stress:	stressed, unstressed
v/uv flag:	voiced, unvoiced
place:	bilabial, alveolar, palatal, velar
manner:	plosive, nasal, fricative, approximant
For consonant:	
constriction position:	front, central, back
constriction level:	close, (cloase-mid, open-mid), open
consonant place:	bilabial, alveolar, palatal, velar
consonant manner:	plosive, nasal, fricative, approximant

flag, place and manner of preceding and following phonemes. For the regression tree of consonant, the constriction position and level of vowel and place and manner of consonant are used. All can take on the value n/a(not applicable) if they do not apply; that is when a vowel is being represented, consonant manner and place are assigned n/a. In this way, every segment is decomposed into multi-valued features that have acceptable complexity for the classification scheme and that have some phonetic justification. These broad classes are described in Table 2.

For the split of each node, all the factors and their corresponding levels will be checked to find a optimum split. And the best split among all the terminal nodes will also be decided, that is for each partition, the most efficient node to split will be choosen and all possible binary partitions of this node have been considered.

Stopping rules:

We need stopping criteria to decide two things, one is the stopping rule for node, the other is the tree size, when to stop the tree growing process. There are some criteria often being used to decide whether a node could be split or not: (1) if the node deviance is less than some small fraction of the root node deviance, (2) if the population of node is smaller than some absolute minimum size. In our system, we adopted the latter criterion. As for the tree size, pruning and shrinking methods can be used to create a good tree, but the tree sequences produced by these methods provide little guidance on what size tree is adequate. The reason is the same data that we used to grow the tree are being asked to provide this additional information. Since the tree was optimized for the supplied data, the tree sequences have no possible alternative but to behave as observed. There are two ways could be used to find the reasonable tree size: one is to use new (independent) data to guide the selection of the right size tree that has been widely used in speech recognition, but for smaller sample sizes, another method that adopted in this paper, called V-fold cross-validation, is preferred. If we examine the cross-validated deviance as a function of tree length, we always can find a minimum is reached for a particular tree length, further increasing the tree length will increase the cross-validated deviance. This is so called over-fitting problem, selection of tree length deserves more attentions.

Prediction:

Least square regression is used for regression tree analysis, the measure of accuracy classifically used in regression is the averaged squared error. And the prediction which minimizes these error is the mean value of the training samples at that node.

3.3 steps

The tree-growing algorithm is as follows:

```
step 1: partition number p = 1, terminal node number tn = 1, only one
        terminal node, the whole space.
```

```
step 2: find an best split among the terminal nodes
```

$$\min_k [\sum_{n \neq k} d_{bs}(n) + d_{as}(k)] \qquad (1 \leq k, n \leq tn) \tag{1}$$

$$d_{as}(k) = \min_{f,l} [d_{as,f,l,1}(k) + d_{as,f,l,2}(k)] \tag{2}$$

where$d_{bs}(k)$ is the total distance value before a split at node k, and$d_{as}(k)$ is the total distance value after a split at node k. f is a number between 1 and m (the number of considered factors), If the factor f has j levels, then l is a number between 1 and $2^{(j-1)} - 1$

```
step 3: choose the optimal value k,f,l, then split node k to 2 nodes
and update terminal node list data, p++, tn++.
```

$$d_{bs}(k) = d_{as,f,l,1}(k) \tag{3}$$

$$d_{bs}(tn) = d_{as,f,l,2}(k) \tag{4}$$

```
step 4: if p is large enough, stop. Otherwise, go to step 2.
```

In prediction, we adopt the convention that a prediction may reside in a nonterminal node if, in following along the path defined by the set of predictor variables for a new observation, a value of a predictor is encountered that has never been seen at that node in the tree-growing process. This is less affected by nonresponse bias, catering better for unknown data type.

Table 3: The statistics of 5 basic vowels used in analysis

vowel	a	e	i	o	u
token no.	148	191	495	133	22
pattern no.	104	119	189	84	17
1-token pattern	76	79	98	57	15
2-token pattern	19	26	34	15	1
3-token pattern	4	10	18	5	0
4-token pattern	3	1	17	5	0
5-token pattern	2	2	4	1	1
>5-token pattern	0	1	18	1	0

4 Evaluations

The tree-growing methods described above were used to generate regression trees for all phone types. Table 3 shows the statistics of 5 basic vowels.

Figure 2 shows that the deviance value always decrease when the partition number increase. Because at each split, the tree algorithm tries to slice the data so as to minimize the inpurity of the descendants. Therefore, the deviance value tends to decrease until only one sample in each node or the tree can not be splitted any more.

Cross-validation

The evaluation method being often used in recognition is divide the data to two sets, one for training and the other one for testing. For smaller sample sizes, V-fold cross-validation is preferred. The basic idea is to randomly divide the spectrum space to some mutually exclusive subspace, for each subspace, a tree is grown by using the information from the remaining subuspaces, the set held out is then used to evaluate the tree then accumulate the total error vs the partition number. As can be seen from figure 3, the characteristics of these curves are a fairly rapid initial decrease followed by a long, flat valley and then a gradual increase for large partition number. The minimum occurs somewhere in the long, flat valley region, where the distance value is almost constant except for up-down changes well within a small range.

Separability:

In our application, we not only want a classification for any future cases, but also some information about how sure we should be that the case is classified correctly. The separability value given that the case falls into a terminal node provides this information. The definition of separability is as follows:

$$s = E[\frac{\text{average between-node deviance estimation}}{\text{within-node deviance estimation}}] \qquad (5)$$

The distinctive features of testing data are used to trace down the tree. After stopping by one of terminal node, the Euclidian distance from the centroid of this node was calculated; this is within-node estimation value. Next, the average distance from the centroids of all other terminal nodes is calculated to find the between-node estimation. These two measures are accumulated for all data. The ratio represents a measure of insulation of the correct category from other confusing categories. Values must be greater than 1 for

Table 4. The significant splits

vowel	@	a	e	i	o	u	@@	aa	ii	oo	ou	uu
1st	pm	pp	pm	pm	pm	pp	pm	pm	pm	s	pm	pp
2nd	pv	s	pp	pp	pp	fp	fm	pp	pp	pp	pp	pm
3rd	pp	pm	pp	pp	fm	X	pp	pp	fm	pp	pp	pp

conventions: s = stress of current phoneme
pv, fv = v/uv flag of previous and following phoneme
pp, fp = place of previous and following phoneme
pm, fm = manner of previous and following phoneme

Consonant	b	d	dh	f	h	k	l	m	n	p	r	s	t	w
1st	pcm	pcm	pcm	pcm	fvl	fcm	fvp	pcm	pvp	pcm	pvl	pcm	pcm	pcm
2nd	pcp	pcm	pcm	pvp	pcm	pcm	pcm	pvl	pvl	pcp	pcp	pcp	fcp	pcm
3rd	fcp	pcp	pvp	fcm	pcm	pvp	fvl	pvp	pvp	fcp	fvp	pvp	pcm	pcp

conventions: pvp, fvp = vowel constriction position of previous and following phoneme
pvl, fvl = vowel constriction level of previous and following phoneme
pcp, fcp = consonant place of previous and following phoneme
pcm, fcm = consonant manner of previous and following phoneme

good selection. Figure 4 shows that the average separability values for the four most common vowels was about 2. This indicates good selection from amongst other competing candidates. Table 4 shows the first three significant split factors from the sequences of regression trees. It is easy to see that place and manner of the previous segment emerge as the important variables for vowel analysis. It is also interesting to note that stress is included as a high-ranking selection criterion for English.

5 Unit Selection

The final objective of tree regression analysis in our application is for unit selection. Two problems must be answered before using these regression trees.:(1) how to solve the missing value problem, (2) how to use the predicting value at each node.

Tree-based model itself is well suited to handle missing value problems, because all the history of regression analysis have been memorized. In prediction procedure, a set of predictor variables will be examined and defining the response values at the deepest node reached as the prediction. Normally this corresponds to a leaf node of the tree, but we adopt the convention that a prediction may reside in a nonterminal node if, in following along the path defined by the set of predictor variables for a new observation, a value of a predictor is encountered that has never been seen at that node in the tree-growing process. We believe it to be less affected by nonresponse bias, because there are no enough information we can trust to decide which is the good path to continue.

In fact, regression trees are used for unknown context, in the case of known context, the unit with same context as target will be used. This kind of process is favorable to non-uniform unit scheme. We adopted a post-process to sort the member of all the node in each tree by using the distance value from the centroid. The prediction process is very straighforward, just trace down the tree and find some appropriate units from the sorting tables. These sorting table can provide not only the CSD measure but also the typicality of synthesis units.

In the synthesis stage, two pass procedures will be adopted. First, to choose the suitable candidates by comparing the context of target phoneme, if unit with same context can not be found in database. The above sorting table will be used to find some appropriate alternative units. The second pass will use the segment continuity, that is the smoothness between segments, to decide the best unit from the candidates list and find the optimum boundary for each segment simultaneously.

6 Discussion

Figure 5 and figure 6 show the cross-validation deviance plotted as a function of partition number. It can be seen that cross-validation does not function efficiently with small numbers of training data, so results were less successful for some classes of phone e.g., diphthongs with less than 20 tokens in the database. Results for consonants showed that some classes were insensitive to contextual variation, but we take this as a positive finding, indicating that less difficulty will be encountered in the substitution of these classes.

7 Conclusion

CSD estimation has been shown to work well for unit selection in Japanese, but it is less successful for English owing to the data scarcity problem. In this paper, we examine the tree regression analysis using broad categories. The cross-validation results show that it seems to be a promising solution for the data scarcity problem in English speech synthesis. Although this kind of analysis is data-driven, the more data we collect, the better the performance we can get. However, the data are never to be enough, the efficient algorithm is still necessary. Furthermore, this technique itself has strong expandibility, it can readily be applied to other feature set, to other languages, and to other speakers.

In this paper, the broad classes have been adopted as distinctive features in analysis, but the further prosody variations such as amplitude, duration and pitch information have not been investigated. We believed the quality of synthetic speech will be improved when these factors were also been incorporated in selection algorithm.

8 Acknowledgements

I would like to express my thanks to Dr.Kurematsu for his kind support of my stay at ATR. Further thanks are due to all the people of ATR who made my stay in Japan as convenient as possible.

References

[1] N.Iwahashi, N.Kaiki and Y.Sagisaka:"Concatenative synthesis by minimum distortion criteria" Proceedings of ICASSP Vol.II, pp.65-68, 1992

[2] M.D.Riley:"Tree-based modelling of segmental durations" pp.265-273 in "Talking machines" edited by G.Bailly and C.Benoit North-Holland, 1992

[3] L.Breiman, J.H.Friedman, R.A.Olshen, C.J.Stone:"Classification and regression trees" The Wadsworth & Brooks, Pacific Grove, California, 1984

[4] L.A.Clark, D.Pregibon:"Tree-based models" In J.M.Chambers and T.J.Hastie, editors, Statistical Models in S, chapter 9, pp.377-419, Wadsworth & Brooks, Pacific Grove, California, 1992

Table 1: Phones table

short vowels	@	a	e	i	o	u	uh		
long vowels	@@	aa	ii	oo	ou	uu			
diphthongs	ai	au	e@	ei	i@	oi	u@		
stop consonants	p	t	k	b	d	g	jh	ch	
fricative consonants	h	f	v	s	z	sh	zh	th	dh
sonorant consonants	m	n	ng	r	l	w	y		

9 Appendix

This part will describe the modifications we have tried when we want to incorporate non-uniform unit synthesis system to English speech synthesis. Every language has its own characteristics. Although our objective is to find a language independent unit selection algorithm, we still have to pay attention to the differences between languages.

A Description of database

The 200-sentence database, recorded by a British female, is used in our experiments. It includes 8802 segments (3223 vowels, 5414 consonants and 165 elision 'phonemes'). The sampling rate is 12 kHz and frame shift is 5ms. Table 1 shows the phoneme categories in English.

A dictionary was created for English database, including orthography phoneme, the phoneme symbols that the speaker really pronunce, the frame information, stress and boundary prosody information. To search the units in this dictionary has same context (or as similar as possible) as target is a important process. To make the search more efficient and fast, we must create some sorting table to help for search. First by defining the phoneme as enumerate variable, we can use the ordering information. And because the environments we condisered include preceding phoneme and following phoneme, so we create a table sorting by preceding phoneme, current phoneme and following phoneme.

B Synthesizer

In NUU system, the excitations include single pulse and white noise. The synthetic speech is generated by using log magnitude approximation filter. In order to improve the quality of synthetic speech, we adopted some modifications in the synthesizer.

1. In a pitch period, the value of the pitch mark pulse is a positive value, and all other pulses are zero. The modification is to assign a small negative value to other pulses, therefore the mean value of a pitch period equals zero.

2. Implement a median filter to solve the inaccuracy problem of pitch value from the database.

C Boundary search

One of the characteristics in NUU system is the moving boundary search. In this way, the laborious unit boundary tuning is replaced by an automatic adaptive unit concatenating algorithm. By examining the environments of preceding unit and following unit, four kinds of acoustic measurements are used to decide the search area. They are minimum low frequency power, minimum cepstrum distance, minimum energy and maximum spectral change.

The junction cost is calculated based on these four measures and find the optimum boundary for each segment simultaneously. This cost can afford a measure to choose the most appropriate unit. At this moment, only the smoothness of spectrum has been considered in NUU system, because it's difficult to find the measure for minimum energy and maximum spectral change. Therefore the junction cost for minimum cepstrum distance is the cepstrum distance between two segments, and the junction cost for the other three cases are set as zero. This value is confused with the real zero cepstrum distance.

Basically, unit selection process of non-uniform system includes two steps. The first is to find an optimal decomposition of the input phoneme sequence and the second is to find the most appropriate extraction and concatenation conditions for each sequence. The measures used in first part are context spectral distance (CSD) and connectibility, the junction cost mentioned above was used for second part. Therefore, the output candidates of first part have been carefully choosen, (the descriptions of first part will be shown on next section), the measures between all candidates should be only one case. From above explanation, if we want to use moving boundary search, we must pay attention to the environments of the candidates.

The other thing that has not been considered in NUU system is the boundary search about consonant clustering context. This is a very important part for English speech synthesis.

D Unit selection

Connectibility and CSD measure are used for the first part of unit selection. Connectibility is a measure of spectral discontinuity between adjacent segments in source database. It can tell us is this a appropriate unit to cut from database. This value is calculated form all the phonemes at the segment point. In English database, we have problems about the accuracy of segment points, so we modify this process. A searching window was set between segments and a maximum spectrum change within this window was used as connectibility. Higher spectrum value means it's easy to segment, lower spectrum value means the spectrum of these two segments are very close. it's not good to use this segment.

Although unit selection requires optimal left and right contexts for a phone, these cannot always be found in the source database. This is the data scarcity problem. To solve it, we must select a phone from an acoustically similar context as a replacement. In NUU system, the CSD estimation has been used for above unknown context problem. The detailed description can be described from figure A.

E Discussion

The algorithm used in NUU system has been confirmed its validity, however, when we try to apply them to English, which has a larger number of vowels and many more consonantal contexts than Japanese, we find many differences. In this report, we pointed out these differences, but the solutions of some problems havn't been proposed. For example, the moving boundary search between two consonants must be put more considerations. The solution of data scarcity problem is investigated in first part, but not completely. We have got some results from the analyses of vowels, but we need more careful analyses about consonants. We hope the synthetic speech could have the comparable quality with NUU Japanese system after these problems being solved.

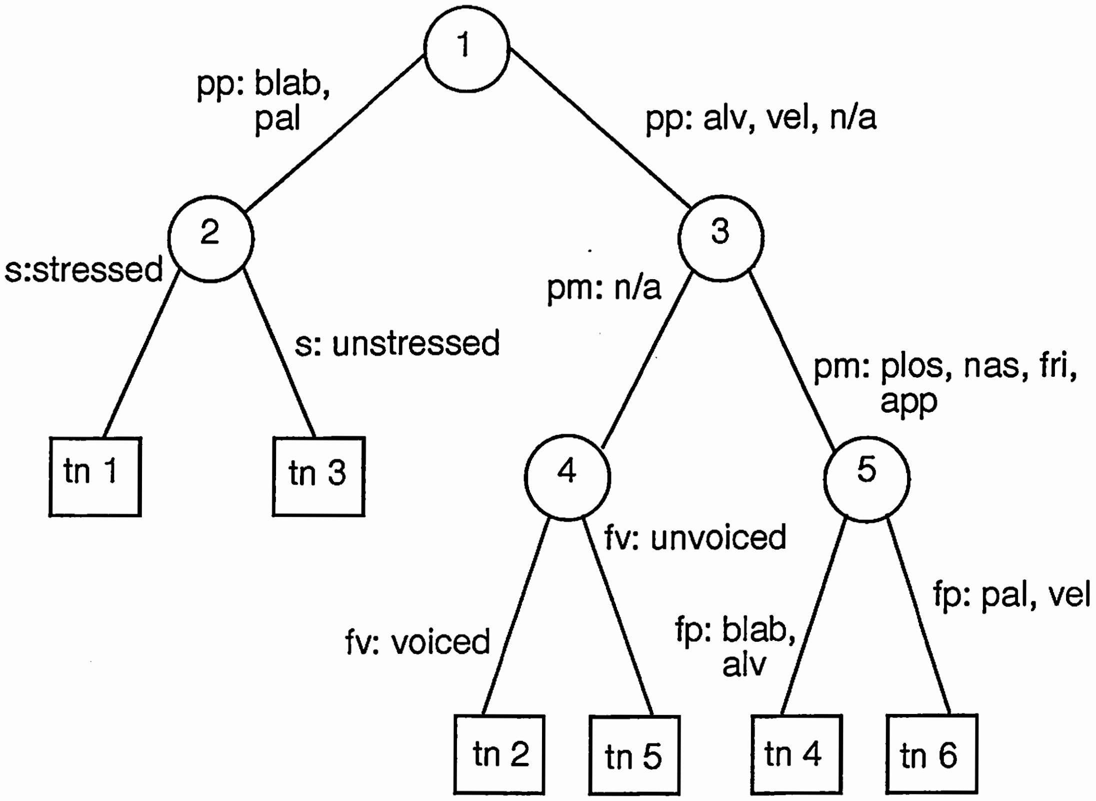

Figure 1. Simplified example of a regression tree for the phone [a]. The circle node labels indicate the split order and the square node labels indicate the terminal node number.

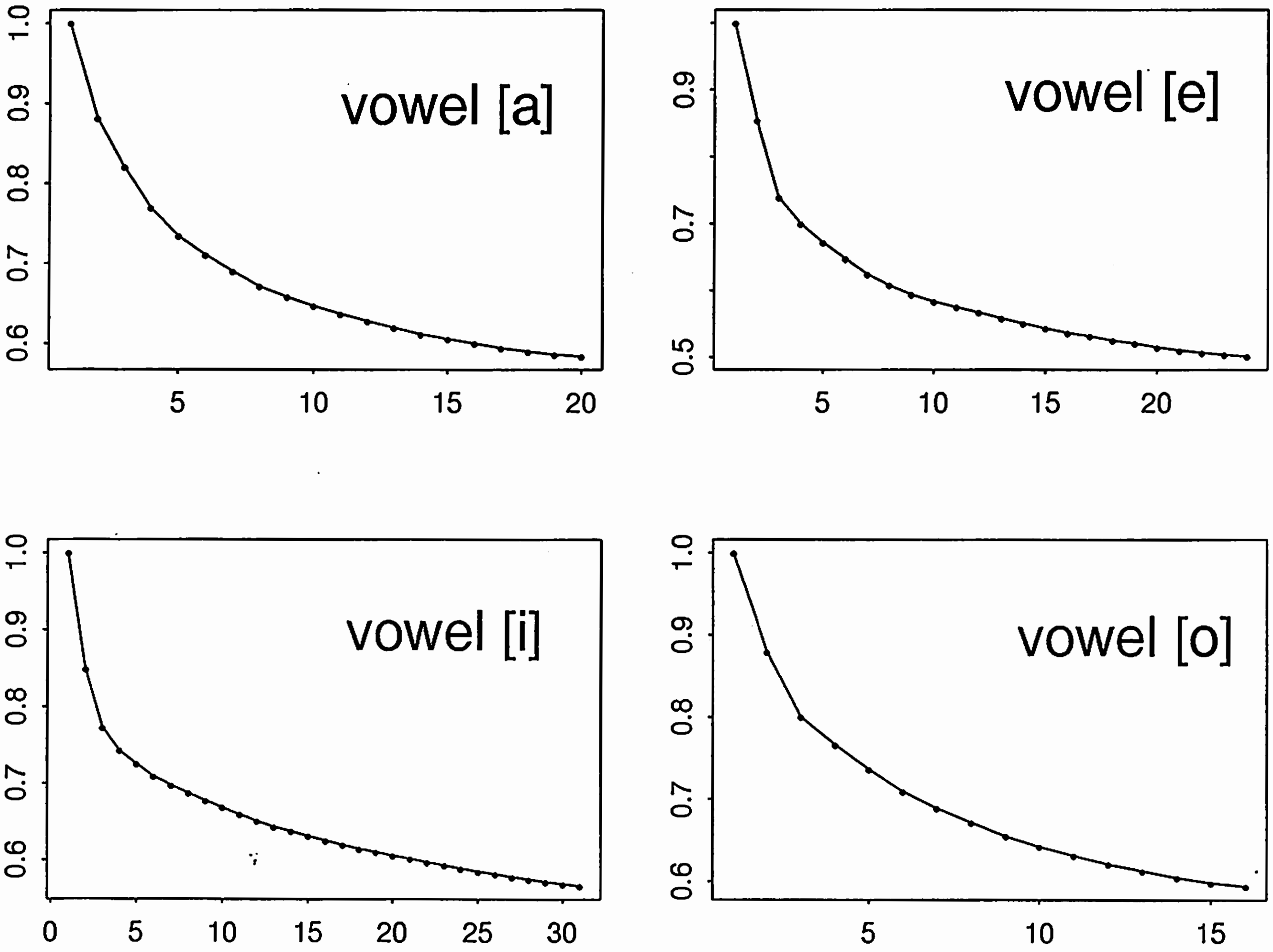

Figure 2: The deviance value of 4 vowels as a function of tree length.

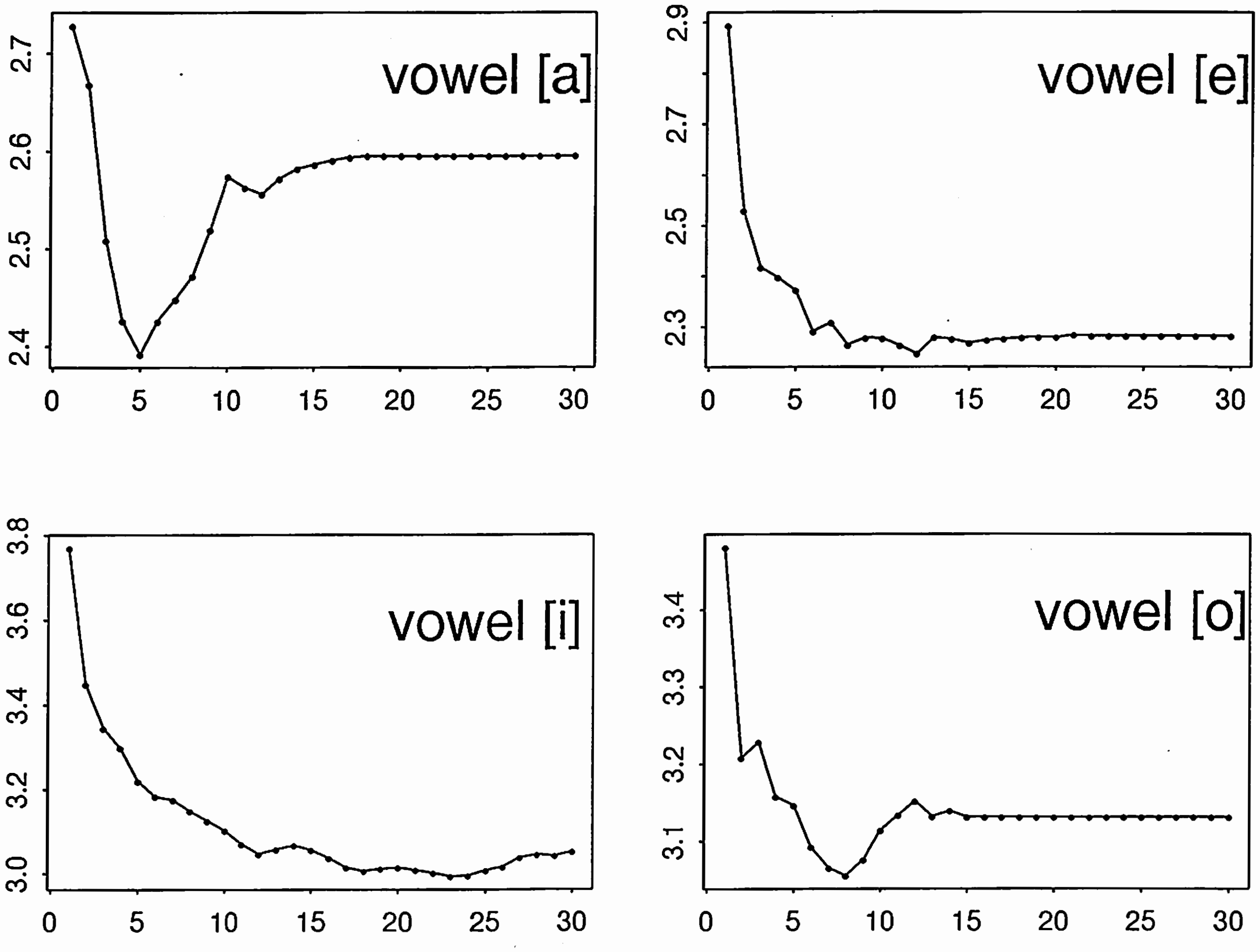

Figure 3: The cross-validation deviance of 4 vowels as a function of tree length.

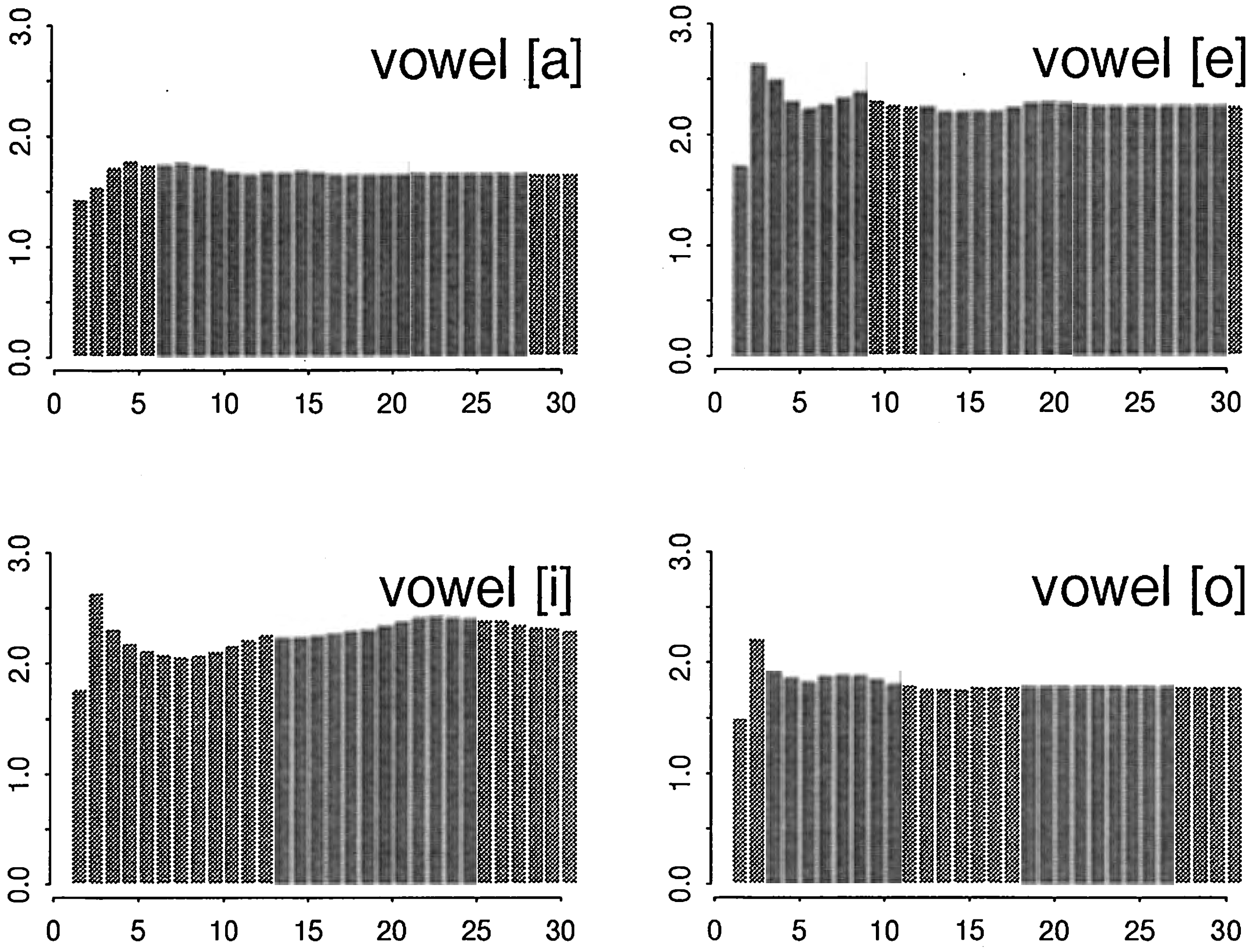

Figure 4: The separability of 4 vowels as a function of tree length.

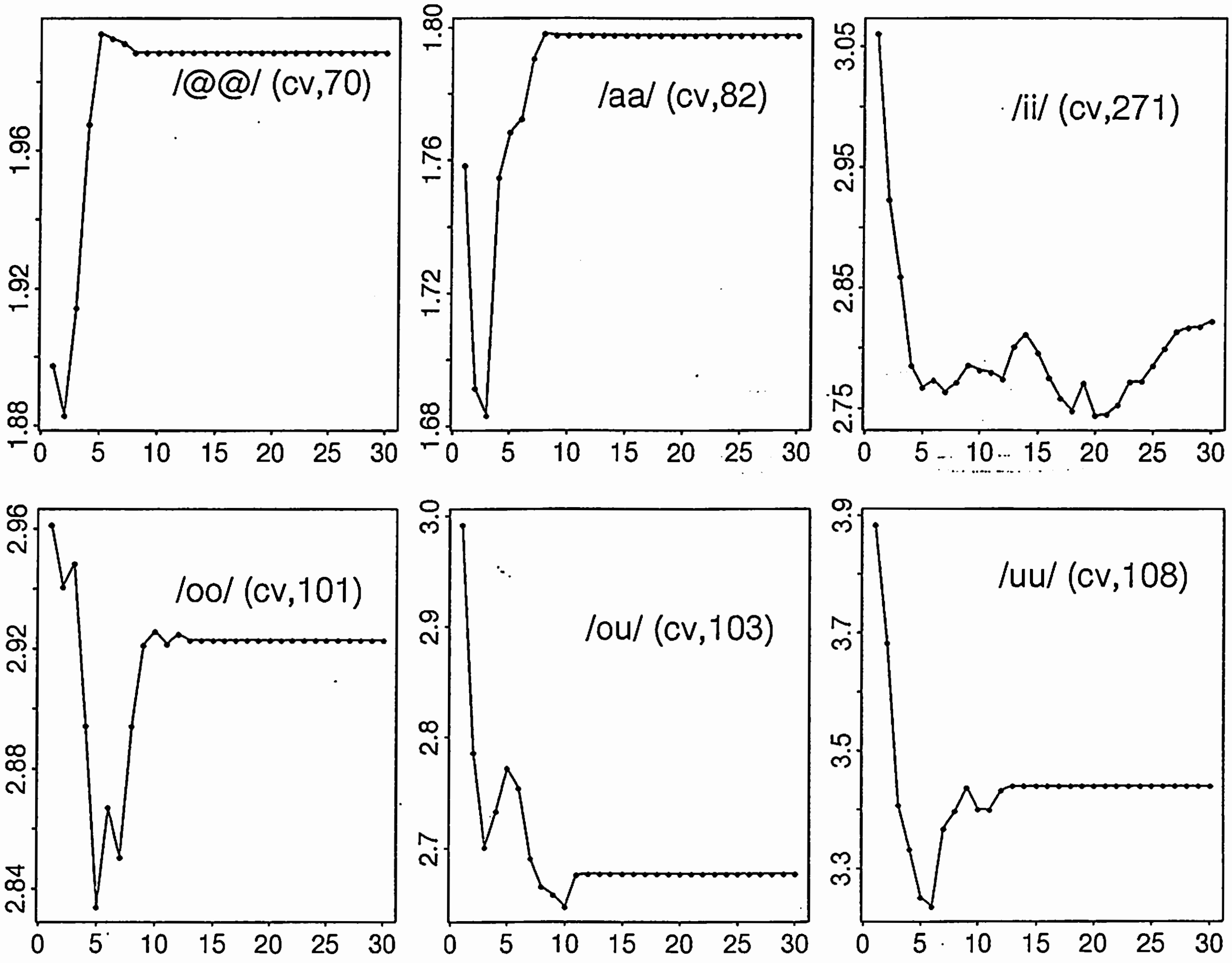

Figure 5: The cross-validation deviance of other vowels as a function of tree length.

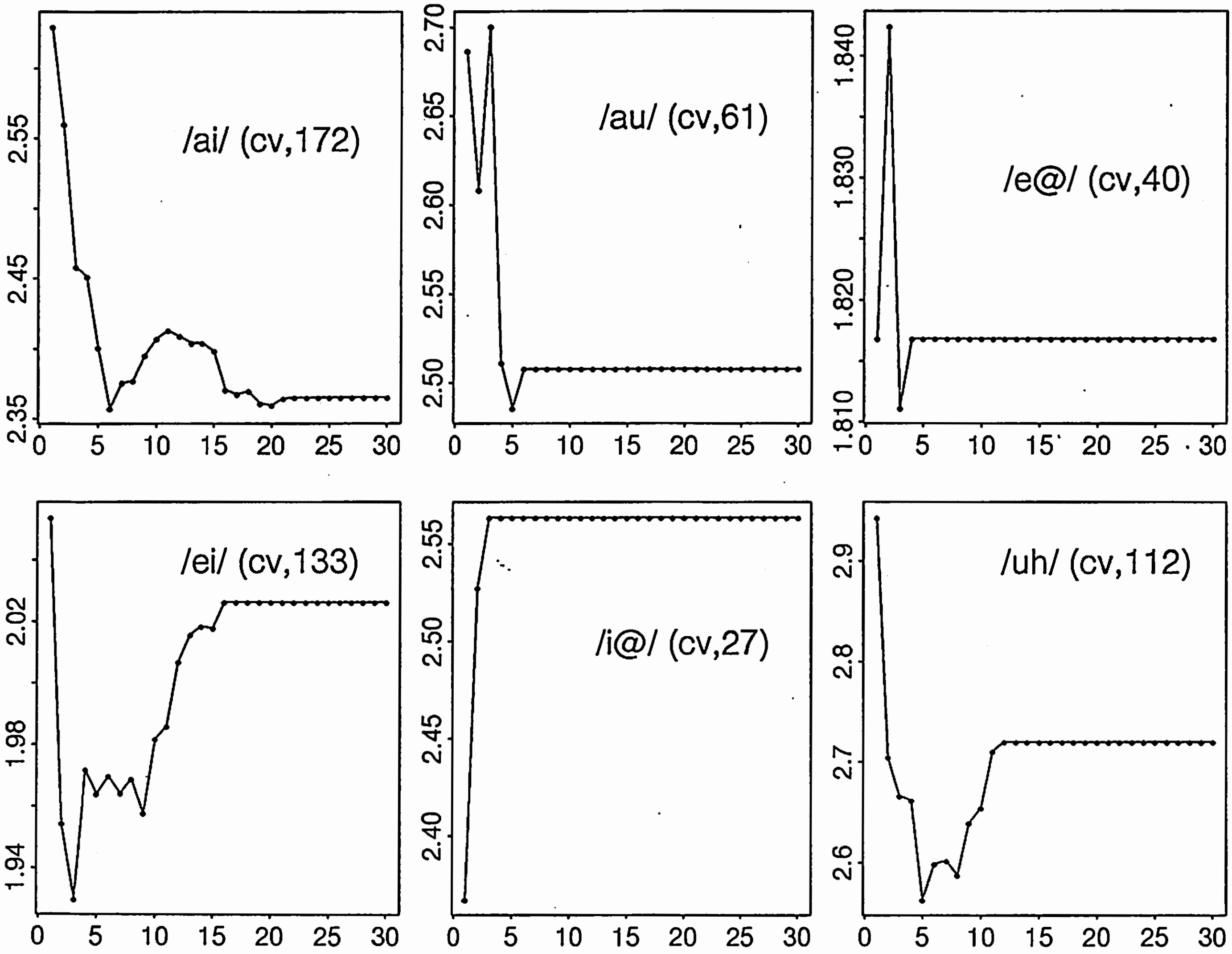

Figure 5: (Cont.)

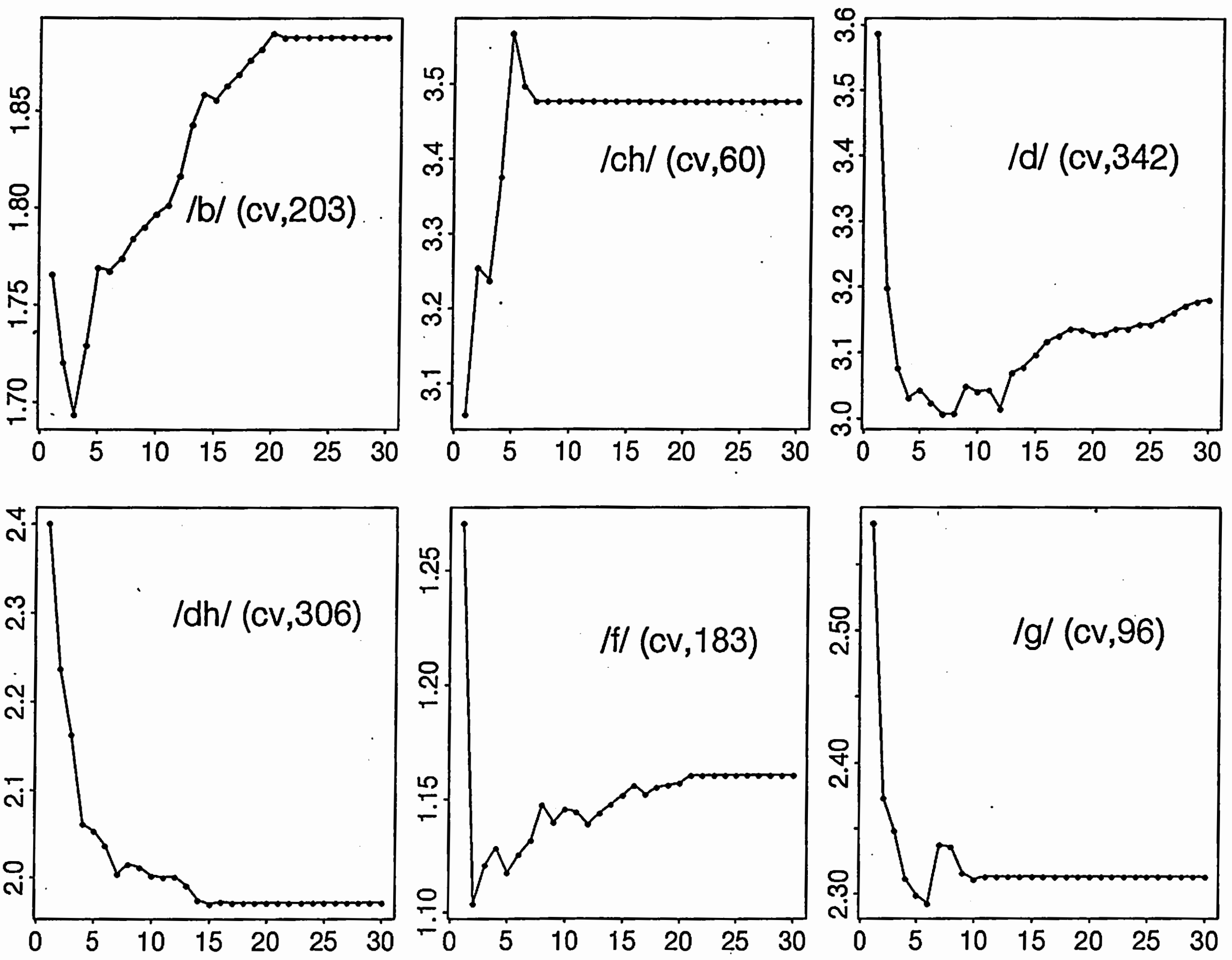

Figure 6: The cross-validation deviance of consonants as a function of tree length.

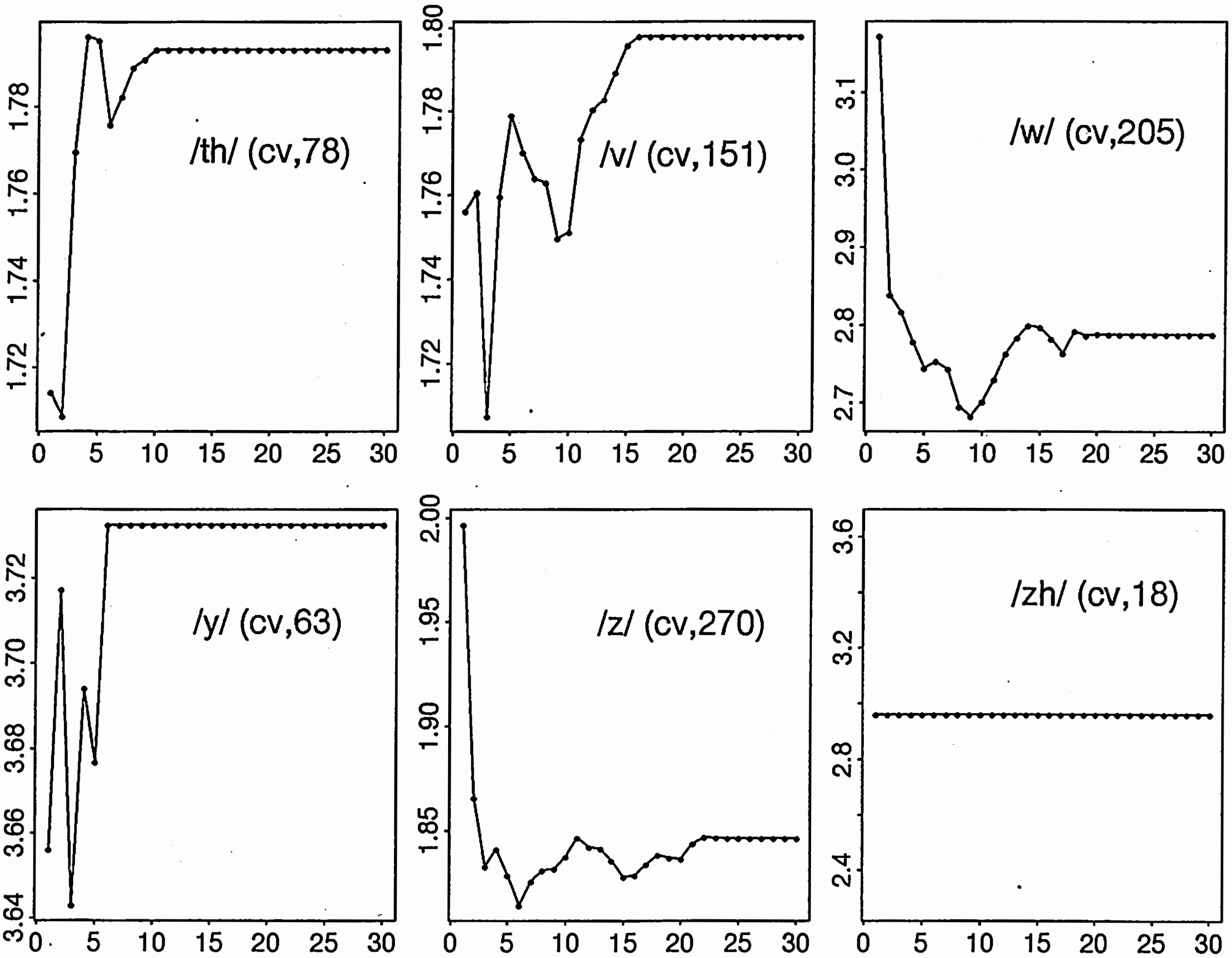

Figure 6: (Cont.)

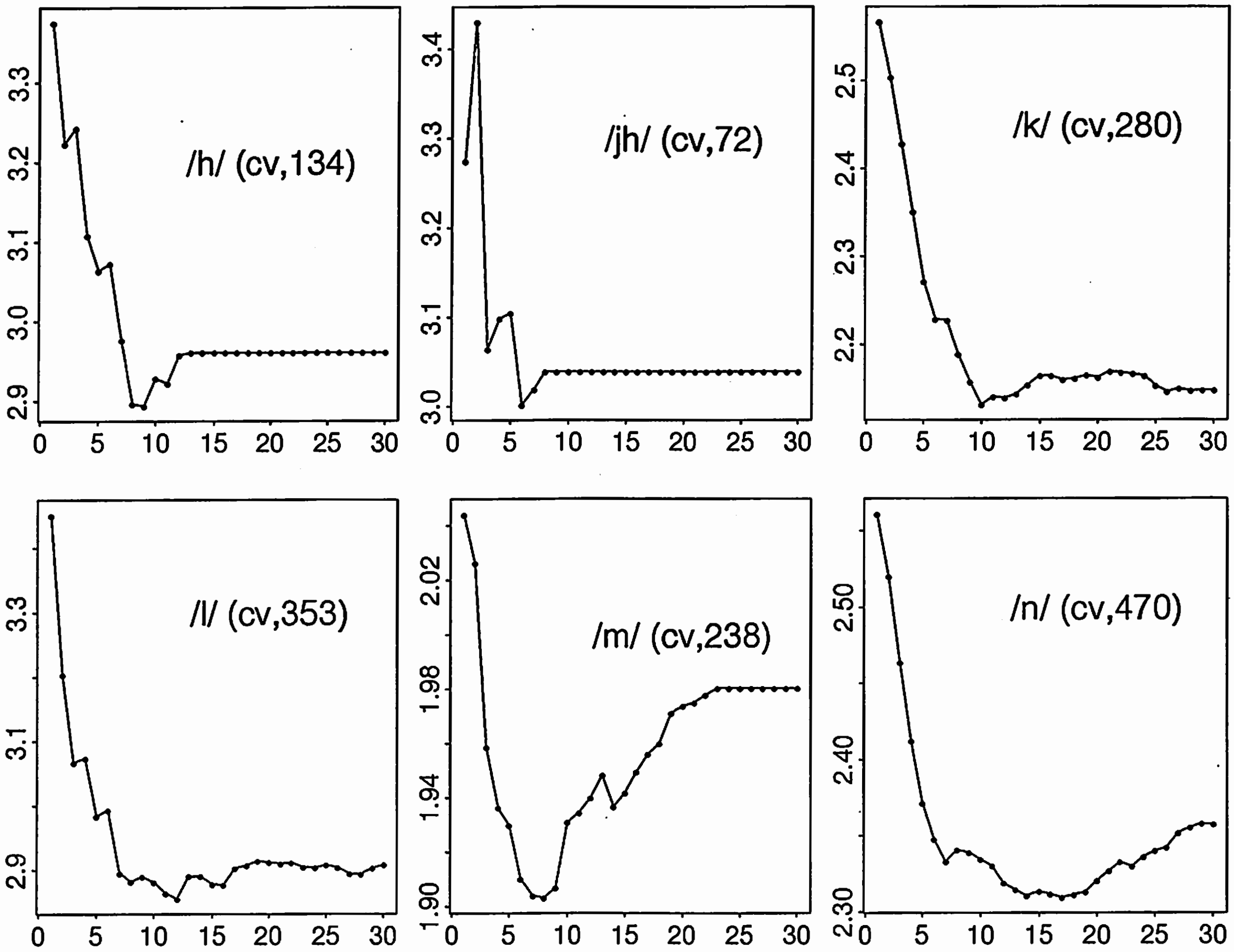

Figure 6: (Cont.)

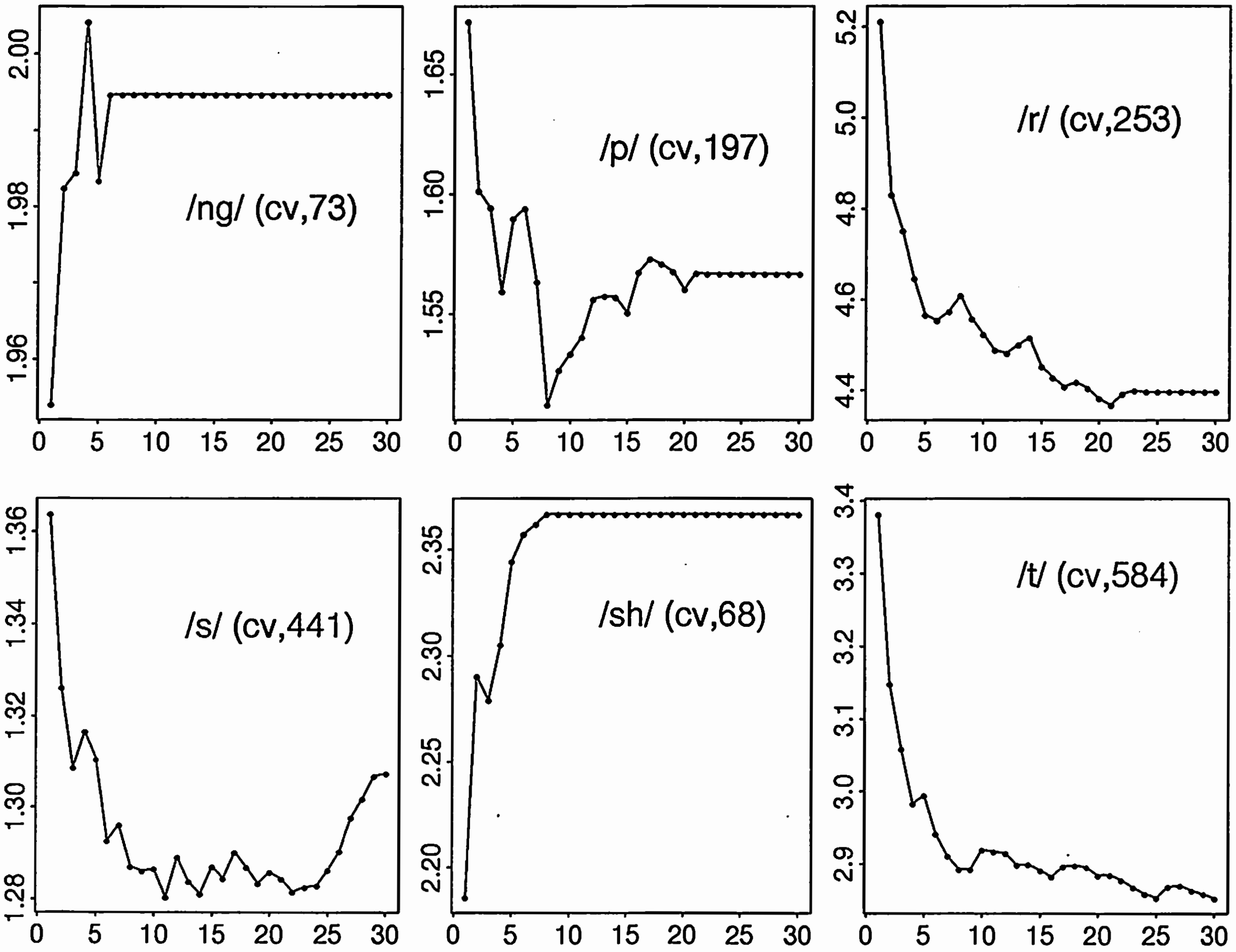

Figure 6: (Cont.)

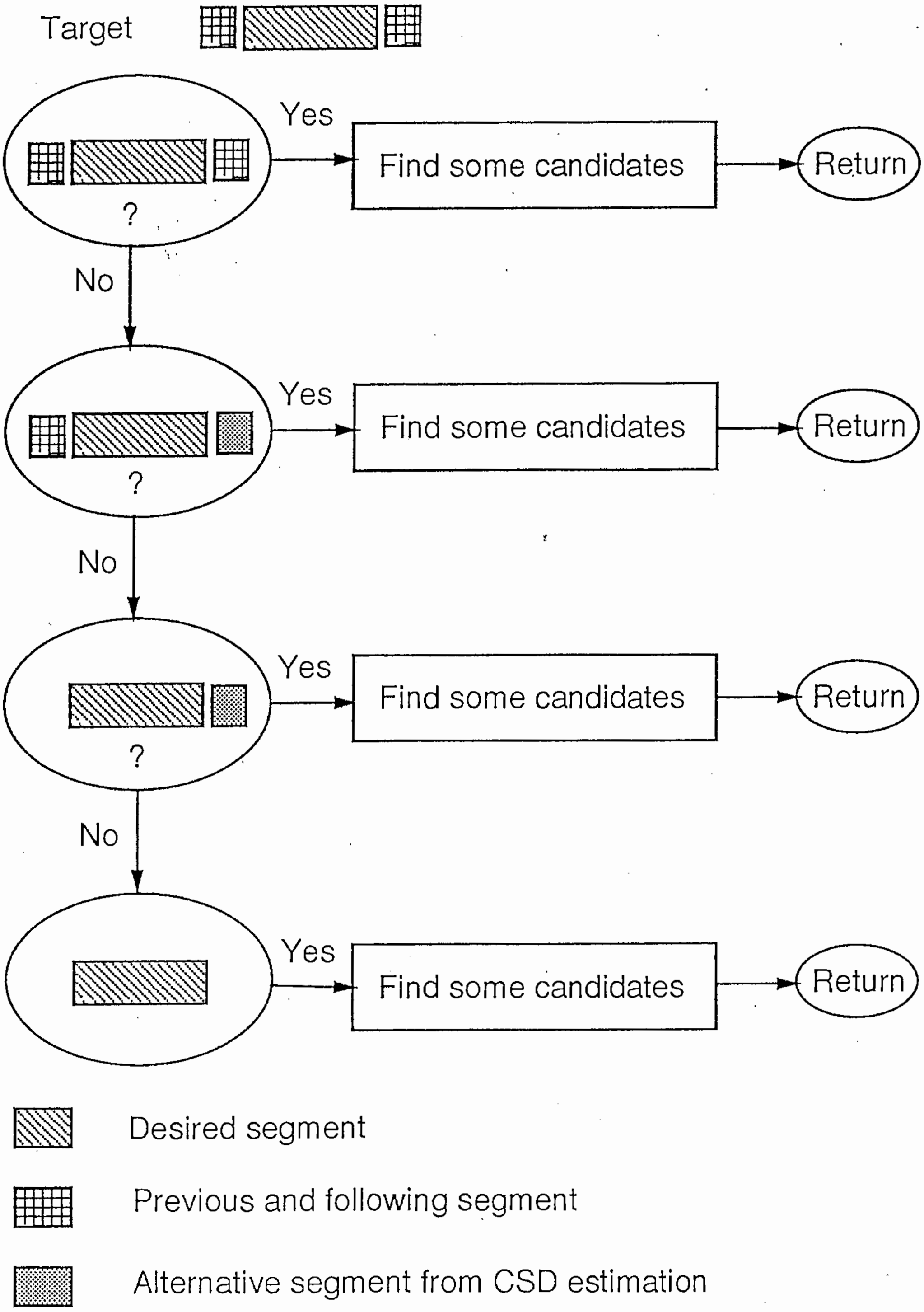

Figure A: The procedures of unit selection.