

## A NEURAL-NETWORK-BASED MODEL OF SEGMENTAL DURATION FOR SPEECH SYNTHESIS

Marcel Riedi

Speech Processing Group, Computer Engineering and Networks Laboratory  
Swiss Federal Institute of Technology (ETH), CH-8092 Zürich, Switzerland  
e-mail: riedi@tik.ee.ethz.ch

### ABSTRACT

This paper presents a neural-network-based model of segmental duration. It was developed with the intention of applying it to speech synthesis for German. Given a set of factors influencing the duration of a phone-sized segment a neural network is used to predict the segment duration. Different mappings of these factors to values suitable for networks with binary and analog input nodes have been applied. So far, the highest correlation coefficient between the observed and predicted segment durations of a test set is 0.886. Informal acoustical tests with this model in combination with a speech synthesis system further demonstrated the feasibility of this approach.

### 1. INTRODUCTION

An important problem in realizing high-quality TTS systems is controlling the prosodic parameters (duration,  $F_0$ , and intensity). Recent approaches to the problem of mapping the phonological information derived from the input text to the parameters necessary for speech production try to learn this relationship automatically (e.g. with a neural network).

The aim of this study was to develop a neural-network-based model of segmental duration which could be integrated into our existing TTS system for German, named SVOX [1]. In this system neural networks have already been successfully used by Traber [2] for  $F_0$  generation.

In experiments on small subsets of Finnish phoneme classes reported by Karjalainen and Alto-saar [3], the phoneme duration was calculated with a neural network. A duration model realized by Campbell [4] successfully uses a neural network to predict syllable durations.

### 2. THE MODEL

The task of our model of segmental duration is to predict the duration of a segment using a set of factors known to affect the duration.

With the exception of plosives the duration of phone-sized segments are predicted. Because of the different effects that some factors have on the hold

and burst part of a plosive, these two parts are treated separately by our model.

From the input text to be synthesized, the SVOX system derives a phonological description consisting of information about phrases, accentuation, syllable and word boundaries as well as sequences of phonemes. This description can be used to determine the values of factors influencing the segmental duration.

The factors chosen for this model are similar to the ones used by Huber [5]. The modifications made were mainly necessary because of differing types of segment used for predicting the duration. The following factors are employed (with their abbreviations given in parentheses for later reference):

*type of the segment (a1,b1, or c1):* The type of the segment is described either by the identity of the segment or distinctive features (depending on the type of neural network input node, see Section 3).

*identity:* The identities of the segments (c1) are used as factor values (e.g. m, ə, r, ...), with 54 different levels.

*distinctive features:* For each segment a pair of factor values differentiate between the possible segment types. For consonants a combination of the type of articulation and a voiced/unvoiced distinction (a1), and the place of articulation (b1) are used as factor values. For vowels a combination of length and position of tongue (a1), and the position of the first formant (b1) are used. Furthermore, diphthongs and preplosive pauses are assigned special factor values. This results in 16 and 13 value levels respectively.

*segmental context (a0,b0,a2,b2, or c0,c2):* The type of the preceding and even more the following segment are known to affect the duration of a segment. These are described by the same factor values as the segment in focus.

*segment position in the syllable (sp):* This factor distinguishes whether the segment is in the onset, nucleus, or coda of the syllable.

*number of segments in the syllable (sn):* The syllables in a sentence have a tendency to be equally

long in duration. This influence on the length of a segment can be taken into account by using the number of segments in a syllable as a factor. For our model four values are distinguished.

*accentuation degree of the syllable (ac):* The accentuation degree of the syllable containing the segment has a strong influence on the duration of the segments, with the main influence being a lengthening of the nucleus. Five different levels are used.

*syllable position in the foot (fp):* It has been shown in [5] that the segment duration depends on the position of the syllable in the foot. Seven different values are used. In German it is not always clear, where to separate two adjacent syllables. In such cases our system shifts the separation to the left. This implies that the accentuation of a syllable can influence the duration of segments of the following syllable. This is taken into account by using the foot position factor, since only the second syllable of a foot has an accentuated syllable preceding it, all others are preceded by an unaccentuated syllable.

*number of syllables in the foot (fn):* Similarly to the effect observed for syllable durations the intervals between two adjacent accentuated syllables (one foot) have a tendency to be equally long in duration. Four factor values are used.

*foot position in the sentence (po):* The lengthening of syllables before sentence pauses and the end of a sentence is taken into account by including the position of the foot within the sentence as factor. The factor values differentiate between sentence initial, sentence final, phrase initial, phrase final, neither initial nor final, and both initial and final (phrases with one foot) position.

### 3. THE NEURAL NETWORK

Figure 1 shows the overall structure of the realized model of segmental duration.

Given a set of values of the factors described in Section 2 the segmental duration is predicted with a neural network. Our model uses a fully connected feed-forward multi-layer perceptron as described, e.g., in [6]. Network parameters that can be varied include the number of hidden layers, and the number of nodes in each layer, all having an influence on the performance of the network. The duration of the segments and the input factors have to be mapped to values suitable for this type of neural network. This mapping is another component affecting the performance of the network. All nodes of the neural networks used in our experiments have a sigmoid logistic activation function  $1/(1 + e^{-u})$ .

The input factors explained in Section 2 have discrete values. An obvious choice for this type of data

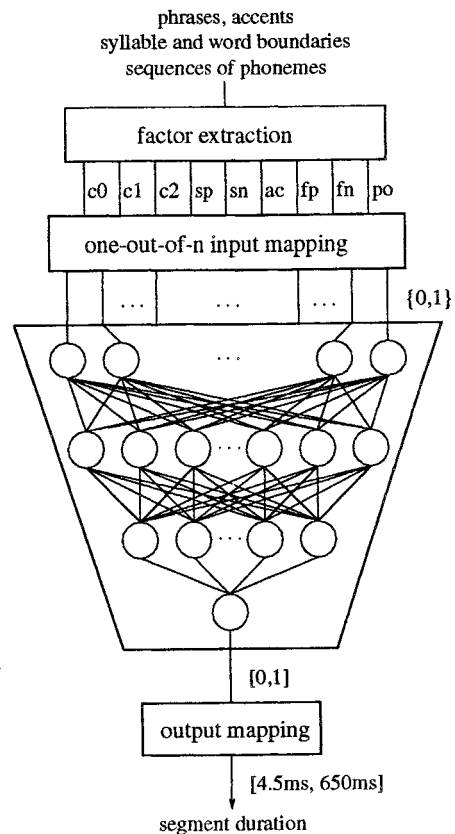


Figure 1: Overall structure of the model of segmental duration. A network with binary input nodes is shown, where each factor with  $n$  values uses  $n$  input nodes for its one-out-of- $n$  mapping.

is a network with binary input nodes. Another alternative are analog input nodes. Combinations of binary and analog units would be another possibility, but wasn't considered in our experiments. The best results were obtained with the following mappings of the inputs and outputs:

*binary inputs:* A one-out-of- $n$  mapping of an input factor with  $n$  values to  $n$  nodes is used, each input node having a value of 0 or 1. The factors  $c0$ ,  $c1$ , and  $c2$  are used with binary inputs (a network with this type of inputs is shown in Figure 1).

*analog inputs:* With analog input nodes the  $n$  values of each factor are mapped to the values  $0, \frac{1}{n-1}, \frac{2}{n-1}, \dots, \frac{n-1}{n-1}$  of an input. The order of this mapping is determined by using the database of segment durations with factor values described in Section 4. For each factor value the mean of the segment duration of all database entries having this factor value is calculated. The higher this mean duration is, the higher the value it is mapped to. To avoid having to map

a factor with a high number of different values ( $c_0$ ,  $c_1$ , and  $c_2$ ) to a single analog input node, the distinctive features ( $a_0$ ,  $a_1$ ,  $a_2$ ,  $b_0$ ,  $b_1$ ,  $b_2$ ) are used for this type of input node.

*analog output:* The duration is a continuously varying value, therefore the different networks trained all have one analog output. The segment duration data (see Section 4) has a log-normal-like distribution with a mean of  $65\text{ ms}$  and a lower and upper bound of  $4.5\text{ ms}$  and  $358\text{ ms}$ . A logarithmic transformation of the data results in a normal-like distribution that can be linearly mapped into the interval  $[0, 1]$ . The reverse transformation (linear mapping and exponentiation) is the mapping used for the network output.

#### 4. TRAINING THE NEURAL NETWORK

The standard back-propagation algorithm [6] was applied to train the neural network.

An important step in using a neural network for duration control is the preparation of suitable samples for training and testing. The speech recordings of 186 sentences from different texts (weather reports, news, tourist information) used for this purpose were spoken by a single trained speaker, an actor (the same corpus was used in [2]).

All sentences were manually segmented. The segments were visually (time-domain speech signal and spectrograms) and acoustically located.

With these segments and information about accentuation, phrasing, syllable boundaries, and phonemes, a database was generated, part of it to be used for training and the rest for testing. Each entry in the database consists of the segmental duration and a set of values of factors as described in Section 2. The database contains 21509 entries.

The factors used for this model result in about  $1.31 \cdot 10^9$  possible factor value combinations, including some combinations that don't exist in natural speech. Having 15849 different combinations in the database there is a missing data percentage of 99.9988%.

By using the same sort of text to generate the database of training and test samples as will occur in an application of the TTS system, the possibility of having equal combinations during training and synthesis increases. However, as can be seen in Figure 2, most of the factor value combinations in our database only appear once.

This shows that an important aspect of a duration model is its ability to predict reasonable durations for combinations not used for training.

The database is divided into two parts, training data and test data (25% for testing). This way a

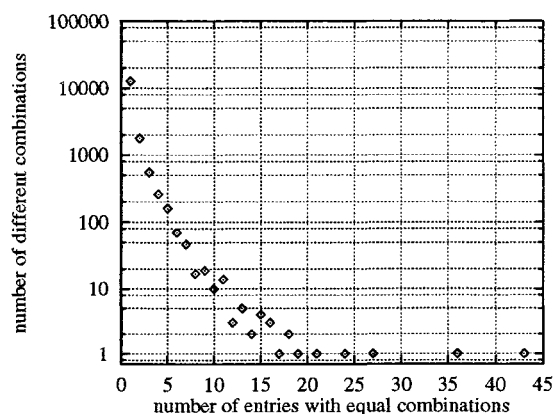


Figure 2: Absolute frequencies of occurrence of factor value combinations in the database (e.g. there are 15 different factor value combinations that appear 4 times in the database).

trained network can be tested on its performance in generalizing. This ability depends on different network parameters. Besides the type of the input nodes already mentioned in Section 3 the following parameters were varied during the repeated cycles of training and testing:

*mapping:* See Section 3. Another possibility examined is a network using the factors  $a_0$ ,  $a_1$ ,  $a_2$ ,  $b_0$ ,  $b_1$ , and  $b_2$  with binary input nodes. This combination did not show any improvement in the performance of the networks and was not considered any further.

*number and size of hidden layers:* Networks with one and two hidden layers were trained. With two hidden layers better results were achieved. The sizes of the hidden layers were varied in the range 8 to 364.

*number of training epochs:* The number of training epochs needed strongly depends on the back-propagation learn rate used. Overlearning has to be avoided by checking the network performance on a test set during training.

181 binary nodes or 12 analog nodes were used in the input layer, and one analog node in the output layer.

#### 5. RESULTS

The different networks trained were compared with each other by calculating the correlation coefficient between the observed durations of the test set and the predicted durations.

For both types of input mapping the network with the highest correlation coefficient found so far is shown in Table 1. The network shown on the left

side of Table 1 uses 181 binary input nodes, 150 nodes in a first and 50 nodes in a second hidden layer, and one output node. With the other type of input node the network size is 12 input nodes, 120 and 40 nodes in a first and second hidden layer respectively, and again one output node.

[181,150,50,1]			[12,120,40,1]		
epochs	$r_{train}$	$r_{test}$	epochs	$r_{train}$	$r_{test}$
30	0.916	0.883	400	0.886	0.865
40	0.921	0.886	900	0.904	0.872
80	0.933	0.882	1800	0.908	0.870
160	0.941	0.878	2100	0.910	0.866

Table 1: Correlation coefficient between measured durations and durations predicted by a network of the training set ( $r_{train}$ ) and the test set ( $r_{test}$ ).

The effect of overlearning can be well seen. While the correlation coefficient of the test data has an optimum after 40 and 900 epochs respectively, the correlation between the observed and predicted durations of the training set still increases. Different learn rates were used for the two networks, partly explaining the differing number of epochs necessary.

To test the contribution of each input node independently a method (the simpler variant) proposed by Campbell [4] was used: Alternately set all input nodes of an input factor to zero and calculate the correlation coefficient using the resulting output of the network. An indication of the factors contribution is then given by the decrease of the correlation coefficient when the nodes the factor is mapped to are set to zero. This method was applied to the two networks (after 40 and 900 epochs) presented above, using the test set (see Table 2).

[181,150,50,1]		[12,120,40,1]			
factor	$r_{test}$	factor	$r_{test}$	factor	$r_{test}$
c1	0.623	a1	0.366	b2	0.800
sp	0.709	b1	0.455	a0	0.822
c2	0.845	b0	0.647	sn	0.822
c0	0.866	sp	0.663	ac	0.829
sn	0.872	a2	0.744	fn	0.843
ac	0.872	fp	0.793	po	0.851
fn	0.873				
po	0.874				
fp	0.880				

Table 2: Correlation coefficient with single input factors set to zero.

According to the data of Table 2 the network with binary input nodes is less sensitive to inputs blinded in such a manner, a possible explanation being the high degree of zero-valued input nodes in the

one-out-of- $n$  mapping. As expected, the factors a1, b1, and c1 strongly contribute to the performance of the model. The factor ac, known to be important for duration control, is correlated to the factor fp and therefore setting one of these two factors to zero doesn't decrease  $r_{test}$  significantly.

Since the correlation coefficient is not necessarily a measure for the acoustic quality of a TTS system using the network, informal acoustical tests were performed by using the neural-network-based model for duration control in the SVOX system.

This tests showed that to some extent the correlation coefficient gives an indication of the suitability of the network for speech synthesis in our TTS system. Having two networks with significantly differing correlation coefficients, the one with the higher correlation clearly produced perceptually more appropriate segment durations. On the other hand with networks having a correlation coefficient above 0.86 very small differences in the generated speech existed.

The results show that even though only a small percentage of all possible factor value combinations is in the database used for training, using a neural network for the prediction of segmental duration is a feasible approach.

With both networks shown in Table 1 segment durations of high quality can be generated. The network with binary inputs does not imply any specific ordering of the discrete values of a factor. Its disadvantage is the larger size of the network, more than five times as many weights are needed.

## REFERENCES

- [1] C. Traber, "Syntactic processing and prosody control in the SVOX TTS system for German," in *Eurospeech 93*, pp. 2099–2102, European Speech Communication Association, 1993.
- [2] C. Traber, " $F_0$  generation with a database of natural  $F_0$  patterns and with a neural network," in *Talking Machines: Theories, Models and Designs* (G. Bailly, C. Benoit, and T. Sawallis, eds.), Elsevier North-Holland, 1992.
- [3] M. Karjalainen and T. Altsosaar, "Phoneme duration rules for speech synthesis by neural networks," in *Eurospeech 91*, pp. 633–636, European Speech Communication Association, 1991.
- [4] W. Campbell, "Analog I/O nets for syllable timing," in *Speech Communication*, vol. 9, pp. 57–61, North-Holland, 1990.
- [5] K. Huber, "A statistical model of duration control for speech synthesis," in *Signal Processing V: Theories and Applications* (L. Torres, E. Masgrau, and M. A. Lagunas, eds.), vol. 2, pp. 1127–1130, Elsevier, 1990.
- [6] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, vol. 4, pp. 4–22, April 1987.