PROSODY GENERATION WITH A NEURAL NETWORK: WEIGHING THE IMPORTANCE OF INPUT PARAMETERS

Gerit P. Sonntag, Thomas Portele, Barbara Heuft[1]

Institut für Kommunikationsforschung und Phonetik, Universität Bonn, Germany
[1] now Lernout & Hauspie Speech Products, Ieper, Belgium
sonntag@ikp.uni-bonn.de

ABSTRACT

As an alternative to synthesis-by-rule, the use of neural networks in speech synthesis has been successfully applied to prosody generation, yet it is not known precisely which input parameters are responsible for good results. The approach presented here tries to quantify the contribution of each input parameter. This is done first by comparing the mean errors of networks trained with only one parameter each and by looking at the performance of a group of networks where each lacks one parameter. In a second approach different networks were perceptually evaluated in a pair comparison test with synthesized stimuli.

1. INTRODUCTION

Generating durations with a neural network has been carried out on the syllable level [1] and on the phoneme level [2,3,4]. Neural networks generating F0 contours have been applied to the isolated word level [5], to the pattern level [6] and to whole phrases [7]. All these approaches use information about content and relative position as input parameters. The importance of these factors is self-evident, but for practical use they must be translated into numerical values. For this, one must decide which are the most important parameters, how they are defined and whether they are continuous or discrete.

2. DATABASE

Our database consists of 10699 syllables and 1185 word or phrase boundaries. It features isolated sentences, questions and answers, and short stories. Contrast emphasis and focal accents are also included. The whole corpus was read by three different speakers. Every syllable and boundary was annotated with parametric information about its content, its relative position, description of F0 maximum, etc. They were all perceptually labelled with prominence values [8]. More than 60 parameters per item are in the database, including context information; some of them are binary, some divide into up to eight categories, some are continuous values. In a first run all networks were trained with parameters from only one speaker (the one that also provided the inventory for our time-domain speech synthesis), in a second run we included all three speakers.

3. INPUT PARAMETERS

The following list contains the parameters chosen for the training of the different networks.

Intonational parameters:

'rifa' - flag, phrase generally rising vs. generally falling
'acce' - flag, syllable with full vowel vs. unaccentable vowel
'hilo' - intrinsic height of the nucleus [0-2]

Durational parameters:

'sonr' - number of sounds within the syllable [0-6]
'foso' - number of sounds following the nucleus [0-2]
'nuctyp' - intrinsic duration of the nucleus [0-4]

Positional parameters:

'pdiss' - distance to preceding accented syllable [0-15]
'fdiss' - distance to following accentuated syllable [0-15]
'bodi' - distance to following phrase boundary [0-25]
'sylnr' - distance to preceding phrase boundary [0-26]
'beatp' - position within foot [0-15]
'beatc' - number of syllables within foot [0-15]
'silgre' - flag, syllable vs. boundary
'type' - grammatical type of phrase [1-8]

Perceptual parameter:

'acat' - prominence value [0-31]

These parameters were chosen from the database as possible input parameters. In some cases their range was compressed to a smaller range, for example the influence of a neighbouring accent syllable is assumed to be only important to a limited extend.

4. OUTPUT PARAMETERS

F0 contours have been automatically parametrized [9] using a maximum based description (see fig.1). The four parameters that describe an F0 maximum are: a) position relative to syllable nucleus, b) height, c) steepness of incline and d) decline. These are the output parameters of the networks for the generation of the F0 contour. The only output of the networks for duration is the syllable length. With this value segmental duration is generated according to the elasticity of the different segments [10].

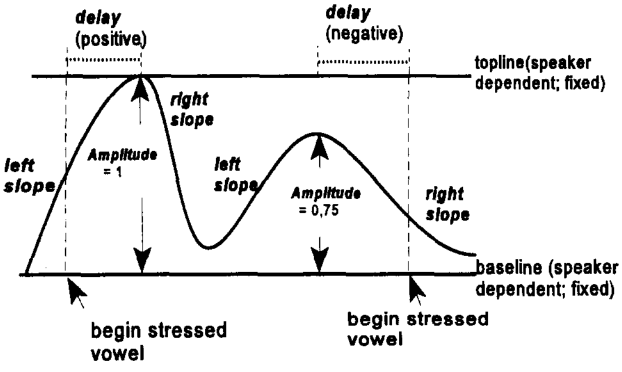

Figure 1: Maximum - based description of F0-contours. Each F0-maximum is described by four parameters.

5. Network Stucture

All networks had a fully-connected feed-forward structure and were trained using the standard backpropagation method. Each network has one hidden layer that contains the same amount of units as the input layer. The number of input units varied according to the choice of input parameters between 1 and 17. The networks for syllable duration have one output unit, representing syllable duration. Here, target output lies in the range of 60-500 ms. The intonation networks have four output units for the model parameters described in section 4. The target values for the 'delay' parameter ranged between -200 and 200ms; according to the model, the other three parameters are relative to fixed top- and baselines. All parameter ranges, input and output, were transformed to fit the [0-1] range.

Because the networks were supposed to determine only the *shape* of the F0 maxima (as the position is implicitly determined by the 'delay' parameter of the model) the input for the intonation networks consisted only of those syllables or boundaries of the database which are associated with a linguistically relevant F0 maximum. Thus, the amount of input patterns for the intonation networks was smaller (4118 when all three speakers were included, 1216 for one speaker) than the amount for the duration networks (11884 for all three, 3730 for one speaker). These input patterns were subdivided into a test set (20%) and a training set (80%).

6. Evaluation by Statistics

The statistical criteria used here to decide the best network performance is the mean output error for the test set. The networks were trained until their performance on the test set started to decline. The number of training epochs needed to reach this state ranges between 100 and 14000.

6.1. Networks with a Single Input Parameter

The simplest way to evaluate the influences of different input parameters on the performance of a network is to train it with each single parameter in turns. Of course, this method neglects all combinatory effects and can thus only be seen as a preliminary approach. The order of input parameters from lowest mean error on the test set to highest mean error is for the intonation network: rifa, acat, acce, silgre, hilo, sylnr, fdiss, pdiss, body, type. The parameter 'rifa' yields a mean error rate of 0,19, the other parameters produce error rates between 0,20 and 0,21. The order for the duration network is: bodi, sonr, beatc, acat, beatp, sylnr, nuctyp, foso, silgre. Here the 'bodi' parameter yields a mean error of 0,03, which translates into 15 ms, the others lie between 0,051 and 0,059.

6.2. Networks with One Missing Input Parameter out of Ten

To determine the usefulness of each parameter with respect to their mutual influence a fixed set of input parameters was chosen. Ten different networks were trained, each with nine of ten input parameters. This method of omitting a different input parameter for each of the networks was thought to be superior to simply setting each of the parameters in turn to zero. Finally we also included one network trained with all ten parameters. The results of the intonation networks suggest that there is only one parameter whose importance seems to stand out from the others. When the parameter 'fdiss' is left out the mean error level on the test set is 0,18. When one of the other parameters is left out the mean error range lies between 0,161 and 0,169. What is pretty surprising is that the performance of the network with all ten parameters yields the poorest result of 0,2 mean error. From that we can only conclude that there exists a certain limit to the number of input parameters which -when it is exceeded- deteriorates the performance of the network.

For duration output of the network consisting of all ten parameters lies within the mean error range [0,0198-0,0221] of seven other networks of the group. Here we can depict the three parameters 'sonr', 'sylnr' and 'bodi' whose absence yield a slightly higher mean error rate within [0,023-0,024]. Thus we may say that the number of sounds within the syllable and the position relative to the phrase are the most important parameters for duration, which is not very surprising.

6.3. Networks Grouped According to Input Parameters

To find out whether parameters describing the content of a syllable or whether those describing the syllable's position within the phrase or the foot are more important, we trained networks with either group of parameters. The input parameters for the two type of networks (duration and intonation) are the same as those of network 5 and network 6 of the perceptual experiment (see table 1 and 2). To compare the networks not only by mean error, we calculated the difference of the actual output value of the trained networks and the target values. We found that for the duration network that the position parameters yielded better results than the content parameters (see fig.2).

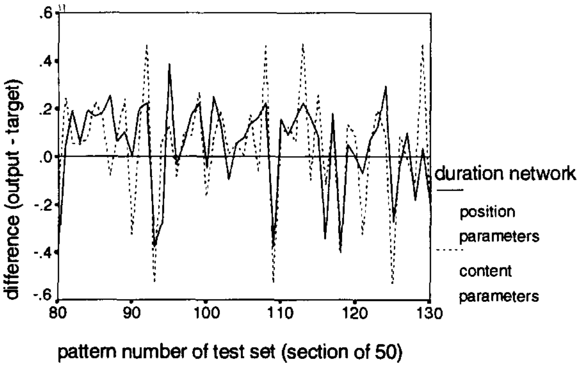

Figure 2: Comparison of two duration networks with different groups of input parameters.

7. Evaluation by Perception

As a mere statistical quantification does probably not say much about the actual performance of a network, a description of a network's ability must comprise some kind of perceptual evaluation. For this task we chose seven trained networks for intonation modelling and duration generation, respectively.

7.1. Stimuli Generation

One long utterance for each type of network ("Die Sonne geht jeden Tag auf, aber sie geht auch wieder unter. Er sagte, er habe keine Zeit." for intonation and "Ich habe siebenundzwanzig rote

net no.	input parameters for the training of duration networks:
1	acat
2	bodi
3	sonr
4	acat bodi
5	beatp beatc silgre foso sylnr bodi (positional parameters)
6	acce acat silgre foso sonr nuctyp (content parameters)
7	beatp beatc silgre foso sylnr bodi acce acat sonr nuctyp

Table 1

Strumpfhosen, zehn Gummibärchen und einen Kühlschrank für meine Großmutter zu Weihnachten gekauft." for duration) was synthesized by our synthesis system [11], generating the prosody parameters with fourteen differently trained networks (see tables 1 and 2). The sentences were chosen due to different types and numbers of segments within syllables and due to different prosodical phrasing, respectively. Input parameters for these networks were generated by the text analysis component. This component determines which syllable will be associated with an F0 maximum. The shape and the relative position of the maximum is then generated by the network.

7.2. TEST DESIGN

Four subjects, experienced with evaluating synthetic speech decided in a simple pair comparison test which stimulus they preferred. They sat in two test runs, one with the different duration networks and one with the different intonation networks. In each test run the seven stimuli were presented together with each other stimulus and in both orders, thus there were 42 pairs to be judged. The stimuli were presented over headphones in a quiet room and the subjects were free to listen to each stimulus pair as many times as they wanted. They were asked to mark the version they preferred and to make use of the notation "equal" only if they were incapable of hearing any differences between the two.

7.3. TEST RESULTS

Looking at all the trained networks, we found the mean range of the output values (measured as standard deviation) to be smaller than the range of the target values for all output parameters (see fig.3). Somehow, all our networks seem to transform the input parameters into more or less average output parameters. This is definitely one of the reasons why we found little differences between them.

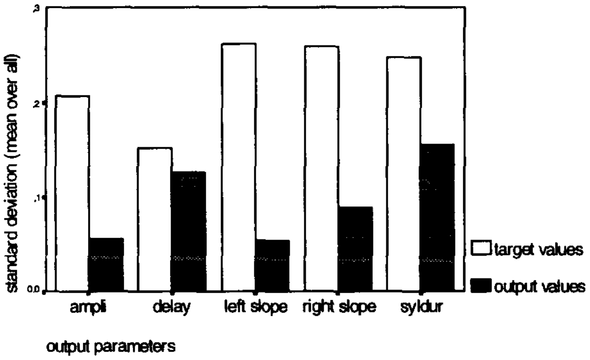

Figure 3: Comparison of the standard deviation of output and target values.

net no.	input parameters for the training of intonation networks:
1	acat
2	acce
3	silgre
4	acat rifa silgre
5	acce rifa silgre hilo acat (content parameters)
6	fdiss pdiss silgre bodi synr rifa (positional parameters)
7	fdiss pdiss silgre bodi sylnr rifa acce acat hilo

Table 2

The intonation networks differed little in their output. As a consequence preferences for the networks hardly reach beyond chance level (see fig.4). Subject two perceived 16 pairs as absolutely different. Even subject three who seemed to hear a difference between all the networks did not judge consistently enough to give one network a clear preference over the others. One subject reported that she changed her strategy when she began to compare only one specific phenomenon within the sentence. This task might have been easier with several but shorter sentences. Yet it is remarkable that the network with the most parameters is judged rather inferior to the others. This is in agreement with the numerical results in 6.2.

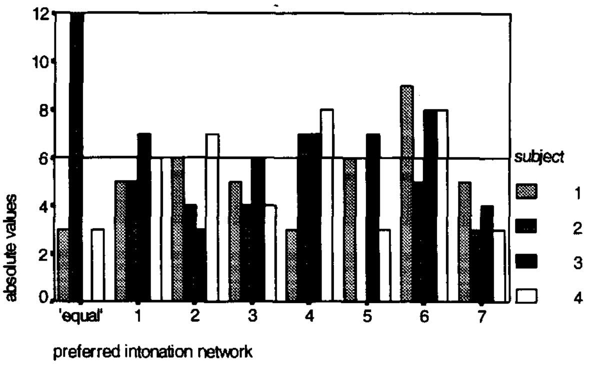

Figure 4: Subjects' preferences of the intonation networks, the reference line indicates chance level.

The differences between the duration networks proved to be perceivable. The obvious favourite of all subjects was the network with the largest number of input parameters (see fig.5). Here only one pair was judged to be equal and the network that had only the perceptual parameter as input came last.

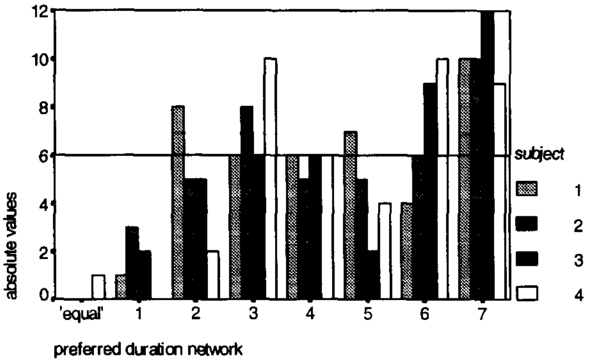

Figure 5: Subjects' preferences of the duration networks, the reference line indicates chance level.

To compare the perceptual judgements to numerical evaluation, we calculated the correlation of target and output values for each output parameter. Fig.6 shows that the output parameter 'delay' was better computed with the positional information of network six. F0 amplitude improved with the content information of network five.

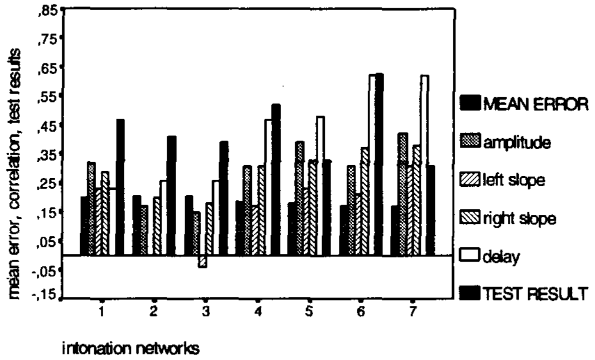

Figure 6: Comparison of mean error, correlation of target and output values and perceptual preference for intonation networks.

Looking at the duration networks it is interesting to find that positional information (network two and five) yields a better numerical result than content information (network three and six). Yet, for the perceptual evaluation content information seems to be more important (see fig.7).

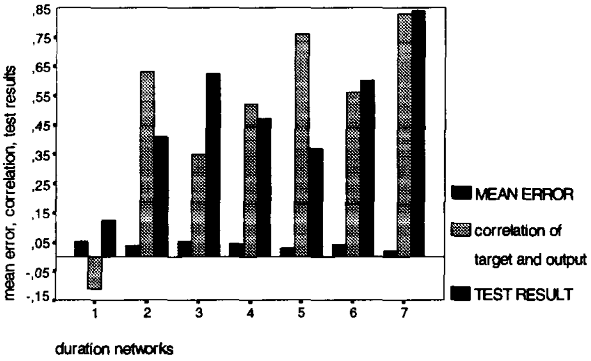

Figure 7: Comparison of mean error, correlation of target and output values and perceptual preference for duration networks.

8. Conclusion

A general observation that we made is that our networks compress the range of the output parameters compared to the range of the target values. Therefore the differences between the synthesized test stimuli were rather small.This may be due to the fact that we did not distinguish between binary and continuous input parameters. Another finding is that there seems to be a limit to the number of input parameters (for our intonation networks) beyond which the networks performance starts to deteriorate. The attempt of grouping the input parameters into positional and content information showed that positional parameters are more important for a duration network's performance. The fact that listeners seem to respond more to content parameters needs to be further investigated.

This work has partly been funded by the German Federal Ministry of Education, Science, Research and Technology in the scope of the VERBMOBIL project under Grant 01 IV 101 G.

References

[1] Campbell, W.N. (1990) "Analog I/O nets for syllable timing" *Speech Communication 9*, pp.57-61

[2] Riedi, M. (1995) "A neural-network-based model of segmental duration for speech synthesis" *Proc. Eurospeech'95*, Madrid, pp.599-602

[3] Karjalainen, M; Altosaar, T. (1991) "Phoneme duration rules for speech synthesis by neural networks" *Proc. Eurospeech'91*, Genova, pp.633-636

[4] Mana, F.; Quazza, S. (1995) "Text-to-speech oriented automatic learning of Italian prosody" *Proc. Eurospeech'95*, Madrid, pp.589-593

[5] Scordilis, M.S.; Gowdy, J.N. (1989) "Neural network based generation of fundamental frequency contours" *Proc. ICASSP*, pp.325-328

[6] Sagisaka, Y. (1990) "On the prediction of global F0 shape for Japanese text-to-speech" *Proc. ICASSP*, pp.325-328

[7] Traber, C. (1992) "F0 generation with a database of natural F0 patterns and with a neural network" in: *Talking Machines; Theory, Models and Designs*, North Holland: Elsevier, pp.287-304

[8] Portele, T.; Heuft, B. (1996): "Towards a prominence-based speech synthesis system" *Speech Communication 20*, (to appear)

[9] Heuft, B.; Portele, T.; Höfer, F.; Krämer, J.; Meyer, H.; Rauth, M.; Sonntag, G. (1995): "Parametric Description of F_0 contours in a Prosodic Database" *Proc. ICPhS 95*, Stockholm, pp.378-381

[10] Campbell,W.N.; Isard, S.D. (1991): "Segment durations in a syllable frame" *Journal of Phonetics 19*, pp.29-38

[11] Portele, T.; Höfer, F.; Hess, W.; (1994): "Structure and Representation of an Inventory for German Speech Synthesis" *Proc. ICSLP*, Yokohama, Japan, pp.1759-1762