BMBF

Verbmobil
Verbundvorhaben

Prosody Generation with a Neural Network

Thomas Portele
Andre Reuter
Barbara Heuft

IKP Universität Bonn

Vm **Report 175**
September 1996

September 1996

Thomas Portele
Andre Reuter
Barbara Heuft

Institut für Kommunikationsforschung und Phonetik
Universität Bonn
Poppelsdorfer Allee 47
53115 Bonn

Tel.: (0228) 7356 - 80
Fax: (0228) 7356 - 39
e-mail: {tpo}@ikp.uni-bonn.de

Gehört zum Antragsabschnitt:
4.3

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 D 08 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

Prosody Generation with a Neural Network

Thomas Portele, André Reuter, Barbara Heuft
Institut für Kommunikationsforschung und Phonetik, Universität Bonn
Poppelsdorfer Allee, D-53115 Bonn, Germany
email: tpo@ikp.uni-bonn.de

Abstract: The use of neural networks in speech synthesis has been especially successful in the domain of prosody generation. The approach presented here differs from others in a) the transformation from a simple input to an output vector consisting of different parameters and b) the use of subcorpora that allow specialized networks. The network operates in a prominence-based synthesis system, where prominence is the most important parameter and is, consequently, the input parameter for the network. The output is not yet evaluated formally but the synthetic speech sounds natural and lively.

1 Introduction

Speech synthesis is the process of generating speech out of written text. As written text includes less information *prima facie* than spoken text, like speaker, accentuation, phrasing, emotion etc., this information has to be generated during the synthesis process. There are many interdependencies and hidden relationships between e.g. syntax and prosody or context and pronunciation that are not fully understood by linguists and phoneticians. Statistical methods, however, have been proven to be very effective in these tasks. Especially neural networks are used due to their automatic learning ability and their power to simulate hidden relations.
Campbell (1990) described a neural network to generate syllable durations for English, Karjalainen and Altosaar (1991) one to determine phoneme durations; Traber (1992) used a neural network to generate F_0 contours. After their pioneering work other systems have been described, for instance for Italian phoneme durations (Mana & Quazza, 1995). However, there are inherent difficulties in an unconstrained statistical approach like a neural network (van Santen, 1994).
Our approach differs from the ones described in the literature in two ways. Firstly, our transformation from input to output transforms a simple description into a set of different but mutually depending prosodic parameters, instead of using multiple input information to generate one parameter like F_0 or duration. Secondly, the training set was designed carefully to have not only a wide coverage of well-known phenomena like contrast emphasis but also to use subsets of similar structure as an indication on how good the network is able to perform generalizations.

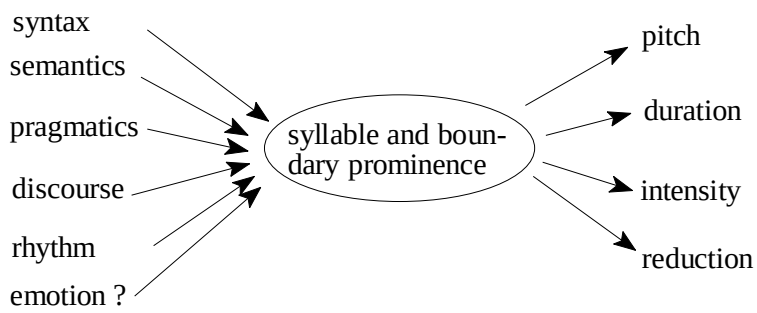

Figure 1 Structure of the prominence-based synthesis system. The transition from prominence to pitch and duration values is carried out by a neural network.

2 Synthesis System

We are currently working on a *prominence-based* synthesis system (Portele & Heuft, 1996). There exists an intermediate description in this system that is very simple and intuitive and, on the other hand, sufficient to capture the relevant phenomena from higher levels and to allow lower level algorithms to transform them into adequate output. This description consists of *content* and *prominence*. Every syllable and every boundary has content information (the phonemes for the syllable, the boundary type [i.e. rising/falling] for the boundaries) and an associated prominence value (between 0 and 31 for syllables, between 0 and 9 for boundaries). Figure 1 displays the structure of such a system. We

have argued for its validity elsewhere (Portele & Heuft, 1996); in this context it is the simplicity of the approach that is important. The network should perform the transformation from prominence and content to pitch and duration information and, eventually, to intensity and level of articulatory reduction as well. However, the current version of our training set does not include information about the latter parameters.
We assume that mismatches between prosodic parameters that can be very disturbing to a listener will be reduced in number and severity compared to the standard method of generating duration and F_0 independently from each other. Instead of using the F_0 values in Hz as output values (Traber, 1992) a method was developed to parametrize an F_0 contour (Portele et al., 1995). Each perceptually relevant F_0 maximum asscociated either with an accented syllable or a rise boundary is described by four numeric values (position, amplitude, left and right slope). An algorithm was implemented that extracts these parameters out of a given F_0 contour; they are part of the training set and also part of the output set of the neural network. The parametrized F_0 contours do not differ perceptually from the original ones. Thus, only liguistically relevant information for accented syllables in terms of phonetically meaningful parameters is learned by the network.

3 Input

The selection of the input set is critical for the success of the network as a generator. Because it is not possible to capture all combinations of potential influences we decided to generate several indepentent sets that can be used as complete training sets or as part of a combined set (Heuft et al., 1995). The use of the network in a face-to-face translation system (Wahlster, 1993) for negotiation dialogues led to three groups of question-answer pairs: yes-no questions, wh-questions, and questions with <or>. Here, contrast emphasis and focal accents are featured. To synthesize messages to the user, some orders like “Speak louder, please!” were grouped in a fourth set. A fifth set consisted of short statements uttered in a neutral tone; a sixth set was composed of short stories. The total corpus consists of over 3600 syllables. It was read by three speakers; one was chosen to become the model speaker for the network because of her very natural and pleasant intonation.
As said before, the prominence-based synthesis system assumes that prominence and content information are sufficient parameters. For syllables the content information was split and coded into three values, namely number of phonemes, type of vowel, number of postvocalic sonorants. The boundaries were excluded and coded in three values: distance from preceding boundary, distance to following boundary, type of following boundary. Thus, seven input values for each syllable were used: three content information parameters, three boundary parameters, and prominence. They were coded into real numbers between -1 and 1. In the current version of the network, a left and a right context of two syllables is used. An input set thus consists of 35 real numbers.

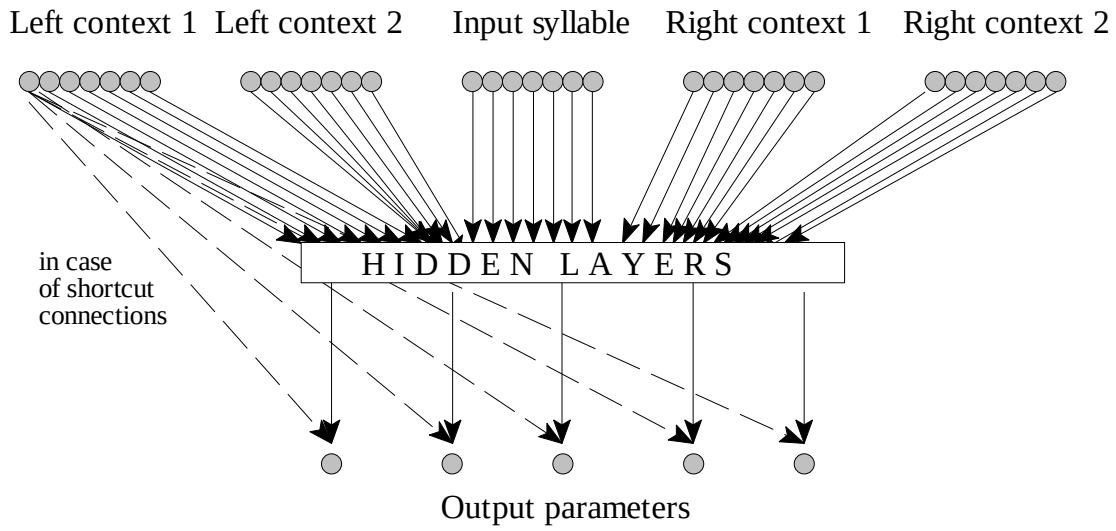

Figure 2 General structure of the network: 35 real input nodes, 5 real output nodes, one or two hidden layers. In case of shortcut connections all the input nodes are directly connected with all the output nodes.

4 Network

The optimal network topology for this task was determined using the general structure displayed in Figure 2. The activation function for each "neuron" was the identity, and the output function was the *tangens hyperbolicus.* Most input and output values were transformed to fit in the [-1,1] interval. These transformations were governed by phonetic knowledge; for instance, the distance to the following boundary could be a value between 0 (directly adjacent) and 25 (far away). The transformation increased the difference between 0 and 1, and decreased the one e.g. between 12 and 13. The appropriateness of these transformations has shown to be a critical factor for the performance of the network.
We compared different versions using the usual feed-forward structure and a feed-forward structure with shortcut connections (Lang & Witbrock, 1989). As a learning algorithm the quickprop algorithm (Fahlman, 1989) was used. The cascade correlation learning algorithm (Fahlman & Lebiere, 1990) and the back propagation algorithm for the feed-forward structure, which were initially also used, were comparably slow and showed no significant increase in performance.
We found that optimal performance can be obtained using a fully connected feed-forward structure with shortcut connections with two hidden layers, the first with 40 and the second with 50 nodes.
Missing context at the beginning or the end of an utterance was simulated using apparently meaningless context information (mostly zeroes). To avoid the persistence of these pseudosyllables a recurrent structure was not implemented, because the recurrence would lead to an undesirable propagation of this information into the input of the network.

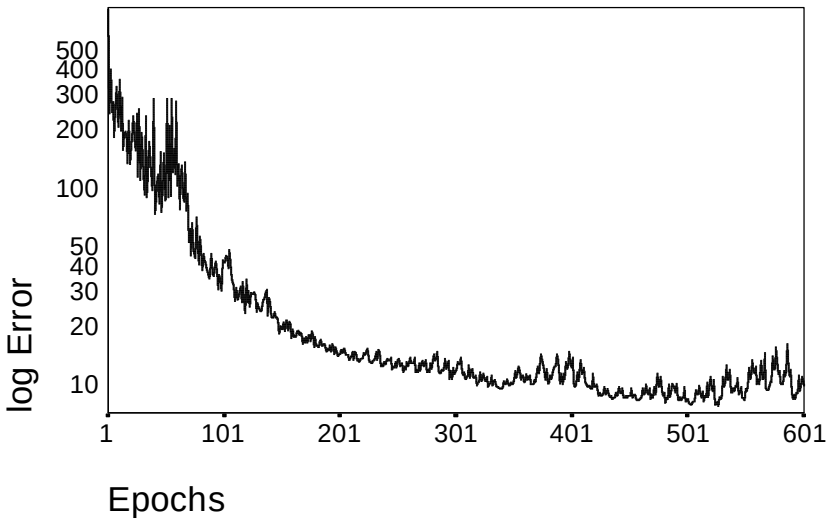

Figure 3
Error function for a typical network.

5 Learning

During the learning process the error decreased not monotonously but steadily up to a certain level. When reaching this point the error goes up and down but stays in the vicinity of this value. Typical smallest error rates are in the area of 4,5 %. Synthetic output produced with networks at different stages of the learning process differed not only in the quality of ist prosody but also in the prosodic phenomena: the phrase-final fall of the F_0 contour was usually not audible before 100 epochs.
Table 1 displays the error distributions for different output parameters for a network with shortcut connections using 80 nodes in one hidden layer. Some parameters are apparently easier to learn than others. Error rates in the magnitude of 10% are tolerable for parameters like syllable duration and F_0 peak height because they are generally below the perception threshold for humans. F_0 peak position can be perceived with more accuracy but the error in predicting it is sufficiently smaller; the network has learned more fine-grained parameters with a corresponding accuracy.

Parameter	error rate %	rate %	Parameter	error rate %	rate %
Syllable duration	≤ 1	10	F0 peak position	≤ 1	47
	≤ 10	82		≤ 3	86
	≤ 20	99		≤ 10	99
F0 peak height	≤ 1	67	F0 slope	≤ 0,1	69
	≤ 5	79		≤ 1	86
	≤ 10	90		≤ 10	99

Table 1 Error distributions for the different output parameters.

6 Results

First results were obtained using the training set consisting of simple statements. When testing the system with statements not in the training set the results were very positive. They were presented during a German prosody workshop and were widely regarded as extraordinary natural sounding and lively. Among the features learned by the network are different levels of prominence, some rhythmic phenomena, and the final fall at the end of a statement. However, more complex structures were not synthesized adequately; this is of course due to their lack in the training set. Currently other training sets are tested if their results for utterances not belonging to the training set but of the same type are as positive as the one for statements. As soon as the training of the subsets is over we will perform a general training with the complete set, and compare the results with those obtained by the specialized sets by a perceptual evaluation.

Acknowledgements
This work was partly funded by the BMBF in the scope of the language & speech project *Verbmobil*, Contract No. 01IV101N0.

References

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

Fahlman, S.E. (1989): “An empirical study of learning speed in back propagation networks.” In: D. Touretsky, G.E. Hinton, T. Sejinowski (Eds.): Proc. Of the 1988 Connectioninst Models Summer School. Morgan Kaufman

Fahlman, S.E.; C.H. Lebiere (1990): “The cascade correlation learning architecture.” In: D. Touretsky (Ed.): Advances in Neural Information Processing Systems 2, Morgan Kaufman, 524-532

Heuft, B.; Portele, T.; Höfer, F.; Krämer, J.; Meyer, H.; Rauth, M.; Sonntag, G. (1995): "Parametric descriptions of F_0-contours in a prosodic database." Proc. XIIIth Int. Congress of Phon. Sci., Stockholm, 2:378-381

Karjalainen, M.; Altosaar, T. (1991): “Phoneme duration rules for speech synthesis by neural networks.” Proc. Eurospeech’91, Genova, 633-636

Lang, K.J.; Witbrock, A.J. (1989): “Learning to tell two spirals apart.” In: D. Touretsky, G.E. Hinton, T. Sejinowski (Eds.): Proc. Of the 1988 Connectioninst Models Summer School. Morgan Kaufman

Mana, F.; Quazza, S. (1995): “Text-to-speech oriented automatic learning of italian prosody.” Proc. Eurospeech’95, Madrid, 589-593

Portele, T.; Krämer, J.; Heuft, B.; Sonntag, G. (1995): "Parametrisierung von Grundfrequenzkonturen." Fortschritte der Akustik - DAGA 95, Saarbrücken, 991-994

Portele, T.; Heuft, B. (1996): “Towards a prominence-based speech synthesis system.” Proc. 2nd SPEAK!-Workshop, Darmstadt

van Santen, J.P. (1994): “Using statistics in text-to-speech system construction.” Proc. 2nd ESCA/IEEE Workshop on Speech Synthesis, New Paltz, 240-243

Traber, C. (1992): "F_0 generation with a database of natural F_0 patterns and with a neural network." In: *Talking machines; Theory, Models, and Designs*. North-Holland: Elsevier, 287-304

Wahlster, W. (1993): “VERBMOBIL: Translation of Face-to-Face Dialogs.” Proc. Eurospeech’93, 29-38