OPTIMIZATION OF A NEURAL NETWORK FOR SPEAKER AND TASK DEPENDENT F_0-GENERATION

Ralf Haury, Martin Holzapfel

Siemens Corp., Otto-Hahn-Ring 6, D-81739 Munich, Germany

ABSTRACT

The generation of a pleasant pitch contour is an important issue for the naturalness of each TTS system. Till now the results are far from being satisfactory. In this paper we present a speaker and task specific approach realized by a neural network. Personal and task specific characteristics are maintained and the demand of generalization decreases. So the results in application can significantly be improved.

Using an optimized network structure global and well localized patterns can be covered and trained simultaneously within one network. Correlation analysis of the data base versus the sensitivity of the trained network validates the importance of distinctive parameters in training. Based on this comparison we give a discussion of the generalization properties of the nn trained speaker and task dependency. Finally a variation of the context range helps to find an optimized tuning of the input parameter set.

1. INTRODUCTION

State of the art TTS systems mainly are highly intelligible but unpleasant to listen to. Unnatural or monotone prosody makes the synthetic speech sounding dull or robotic. Attempts to incorporate more charm in synthetic voices are still at the beginning.

Rule based systems for fundamental frequency contour generation suffer from the lack of a precise and generally agreed on set of rules for to produce a pleasant and human sounding pitch contour. Data driven approaches [7][9] successfully make use of the pitch contours of human voices without explicit formulation of the properties that make a voice sound pleasant. The trainability of those systems makes them open for construction of multilingual systems without the supervision of an human expert.

In section 2 we develop the idea of using the implicitly performed adaption to mimic the prosodic intonation characteristic of one voice in a special task. Section 4 shows details of the parameters that were used as input for neural network (nn). In section 5 we present the architecture for our nn. Section 6 analyses correlations in the training data versus the sensitivities of the trained network and gives a first interpretation of what the network learned in training. In section 7 we report experiments on the variation of the context range used as input parameters.

The presented algorithms for pitch contour generation are part of the new multilingual Siemens TTS-system "Papageno". The prosodic parameters phone, duration and energy are generated from a statistical data base [2].

2. SPEAKER AND TASK DEPENDENT PROSODY

Concatenative speech synthesis uses articulatory tracks taken from utterances of one speaker. Consistently we use a data base of one speaker to train a pitch contour generating network. This network then mimics the prosodic characteristic of this selected voice. The two main advantages of this approach are:

- prosodic characteristics of one voice make the resulting speech sounding much more natural and personal. The impression of being talked to by an androgyn standard voice is reduced.
- Following a single voice is much less complicated training task than the additional demand to generalize over the various speakers in a mixed data base.

Following the idea of specialization we further limit the kind of text to be dealt with. The voice very much depends on concrete situations and moods. A telephone chat with a friend very much differs from a formal speech in public in its intonation and should be treated in a different way.

The task addressed in this paper is reading aloud newspapers. We use a data base recorded with an educated speaker reading approximately 3 hours of text from "Frankfurter Allgemeine Zeitung" containing 1000 declaration sentences of complex structure. We used 70% of the syllables contained for training, 15% for validation, 15% for testing. The data base was automaticaly segmented using a overadapted HMM [2]. A laryngograph was used to accurately detect the moments of glottal excitation.

3. F_0- PARAMETRISATION

Database and synthetic pitch contours are parametrised using a maximum based piecewise linear approximation [4]. Employed parameters are amplitude, delay, left and right slope of an approximating triangle (see figure 1).

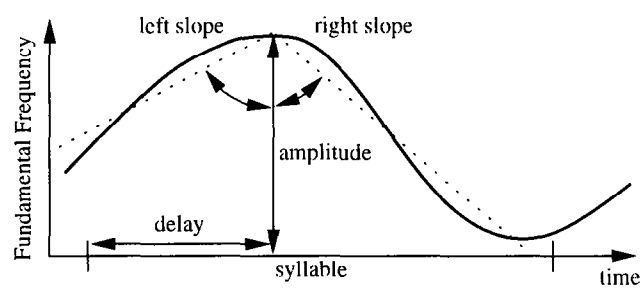

Figure 1: Maximum based description of fundamental frequency

4. INPUT MODELING

Input parameter and pitch contour parametrisation are organized on syllable level. The prosodic marker are generated by the symbolic part of the VERBMOBIL TTS-system [1].

Input parameters are composed of:

1. **Phonetic Information**
   The phonetic information are the phones a syllable consists of. To reduce the number of input variables, we classified the consonants in the four groups liquids, fricatives, orals and nasals.
2. **Prosodic Information**
   - **Prosody Marks**
     This information contains a real valued number denoting the stress of the syllable and flags indicating linguistic features like beginning or end of a sentence [8] (see figure 7 for details).
   - **Durations and Starting Points**
     Additional information is provided relative to the synthesis time axis. The starting point of the syllable, of its vowel and the overall length of the sentence are real valued input parameters.

5. DESIGN OF THE NEURAL NETWORK

5.1. Network Architecture

The Network Architecture is shown in figure 2. Additional to a standard feed-forward multilayer perceptron we use a squaring layer connected to the real valued inputs by an identity matrix (dotted arrow in figure 2). The other matrixes in figure 2 denote a full connection of the captured network nodes or layers. The hidden layer consists of 40 nodes with an activation function tanh(x).

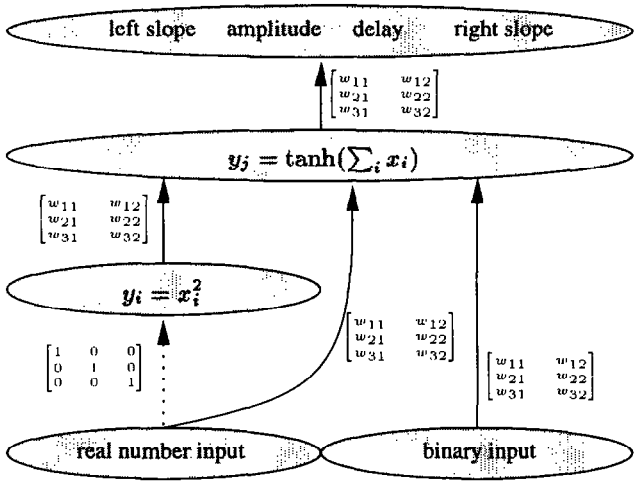

Figure 2: Neural network for F_0-Generation, a full connected feed-forward network with additional squared input variables.

5.2. Linear and Squared Input Parameters

Using linear and squared input parameters easily combines two different classification properties. The direct use of the input parameters results in a linear separation of the feature space. The squared input parameters perform a weighted distance classification by radial basis functions. The first one is capturing global, the second one well localized patterns.

See the activation functions of the hidden layer for both linear and squared inputs in figure 3. Figure 4 shows examples for possible separations of a feature space by both types of classifier.

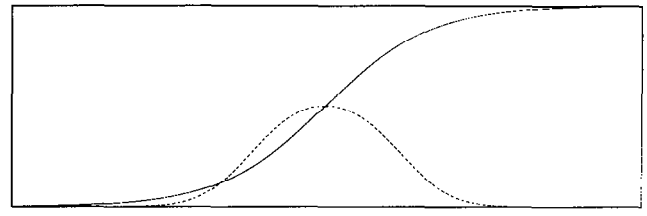

Figure 3: Comparison between $\tanh(x)$ and $\tanh(-x^2)$. The Function $\tanh(x)$ with squared parameter becomes radially symmetric.

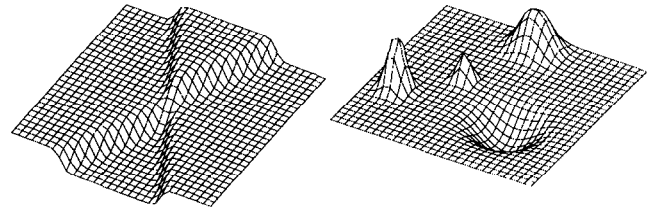

Figure 4: Activation of hidden neurons with two input variables. The left graph shows a linear separation of the feature space by sigmoid functions. The right one shows local activation. Each neuron represents one hill.

5.3. Output Encoding

In first experiments the net showed poor results in learning the output parameters left and right slope. To avoid shortcomings of the squared error function, training these parameters are encoded as a fuzzy set of neurons [5].

Each parameter is represented by the combination of ten neurons denoting equidistant sections of the interval $[0; \frac{\pi}{2}]$. For defuzzification the Center-of-Area-Method showed good results in application.

During training the desired target value of each neuron is defined as a Gaussian centered at the neuron corresponding to the exact parameter taken from the data base (see figure 5).

5.4. Training Strategies

Training starts with randomly initialized weights. In the first training stage we use a gradient algorithm. Vario Eta [6] is based on optimization of the weights by a least mean square criterion of the

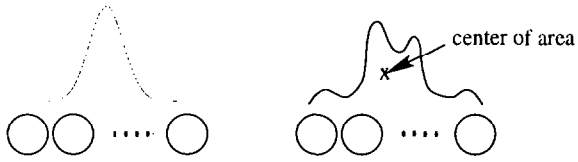

Figure 5: Output encoding for left and right slope. For training we used the Gaussian function. The network output was defuzzificated with Center-of-Area-Method.

error function. The magnitude of weight incrementation in each iteration is varied accordingly to the number of components of the error vector that are reduced by this iteration step. Randomly selecting a subset of training patterns in each iteration step helps to avoid early convergence to local minima.

After reaching an minimum error on the generalization set we change the training algorithm for further refinement in a second stage (figure 6). Now all training patterns are considered simultaneously by an Low-Memory-BFGS method [6]. A modified Newton algorithm is used to minimize the error function described by a second order Taylor approximation.

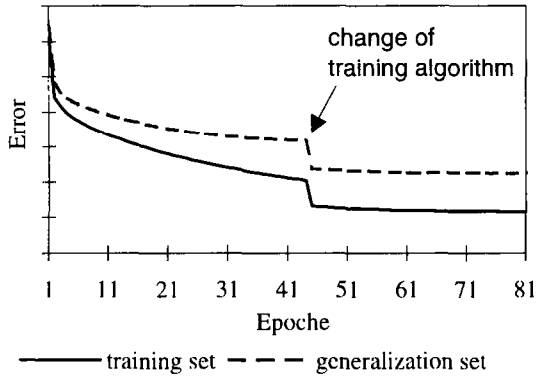

Figure 6: The course of error while training. We changed training algorithm at epoche 43 for further refinement. At the minimum of generalization error we finished the training.

6. SENSITIVITY ANALYSIS

6.1. Analysis Tools

Accordingly to the highly complex combination of input parameters in a nn, analysis of a trained network allows only tendentious interpretation. In this section we discuss the correlation of the training patterns versus the sensitivity of the output nodes defined by their partial derivation

$$\Delta_{ij}(x) = \frac{\partial O_i}{\partial I_j}(x).$$

To get a stable and significant result, the sum of $\Delta_{ij}(x)$ for all patterns of the generalization set was computed as follows:

$$S_{ij} = \sum_x |\Delta_{ij}(x)|$$

As an example we discuss the output parameter "amplitude". Parameter describing the synthesis time axis (see section 4) together with the linguistic flags, do contain redundance. The flag "end of the sentence" e.g. does not provide other information than a value 0.95 for the starting time of the syllable relative to the sentence time axis.

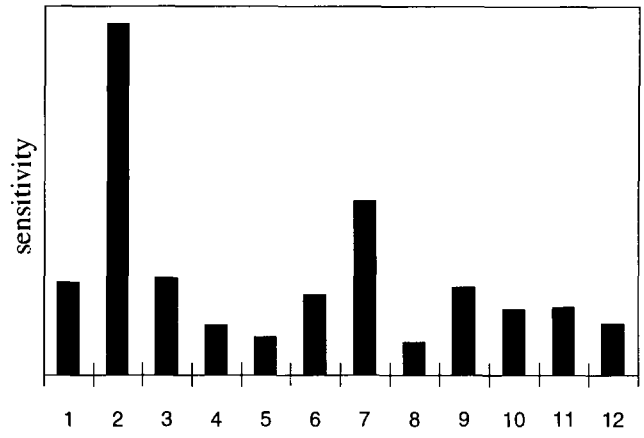

*Figure 7: Sensitivity Analysis of Amplitude. **real number inputs:** 1 duration of sentence, 2 start of syllable, 3 duration of syllable, 4 start of vowel, 5 duration of vowel, 6 stress level, **flags:** 7 beginning of sentence, 8 end of sentence, 9 beginning of extended group, 10 beginning ending group, 11 medium break, 12 short break.*

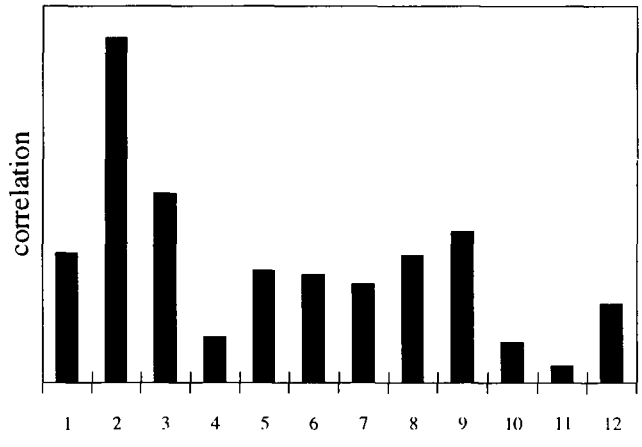

Figure 8: Correlation between Input and Amplitude. For explanation of input variables see Figure 7.

6.2. Discussion

The sensitivities listed in figure 7 are dominated by the influence of the relative position of the syllable in the sentence. The flags

begin and end of sentence (parameter 7 and 8 in figures 7 and 8) are redundant with it. It is interesting to see that the sensitivity to parameter 7 is much higher than to parameter 8, having comparable correlation in the data base (figure 8) and both of them being redundant with the parameter 2.

Parameter 1 captures the overall length of the sentence and so the complexity of its structure. Compared to the input-output correlation the sensitivity of the parameters 9 to 12 is significantly high. These flags denote breaks and the beginning of new phrases. They are not redundant with the overall position of the syllable.

The fact that the data base contains only one type of sentences and the above mentioned structures in correlation and sensitivity encourage the following assumption: "The training results in learning a general macro structure on sentence level. This macro structure is modulated by local influences like phrase boundaries and stress positions."

7. VARIATION OF CONTEXT RANGE

As the pitch contour of natural sentences consist of complex global patterns one would like to take into account a broad context of surrounding syllables. In practical application there are two main constraints:

- Our scope is limited to sentence level. Widely extended context, emphasizes undesirable boundary effects at the beginning and the end of a sentence.
- According to the limited amount of training patterns a more and more complex network lacks of generalization properties.

Table 1 reports a set of experiments varying the context range for both prosodic and phonetic input parameters. Even if the listed generalization error is far from human speech perception it clearly indicates tendencies resulting in audible quality of synthetic speech.

The reported results are behaving in a typical way. Phonetic information mainly is important for the syllable dealt with. Broadening the context of prosodic information improves results up to a certain level. Further extension results in a network with too many degrees of freedom.

For this application we found a clear optimum for a prosodic context of 1 syllable (o*o). Using an extended training data base should enable the training of a network considering a wider prosodic context.

Prosodic Information	Phonemic Information			
	—	*	o*o	oo*oo
—	—	0.1388	0.1328	0.1302
*	0.1144	0.1136	0.1138	0.1148
o*o	0.1109	0.1097	0.1106	0.1119
oo*oo	0.1107	0.1102	0.1105	0.1138

*Table 1: Generalization error of well trained networks dependent on the context width, * one syllable, o*o syllable with a context of one syllable left and right, oo*oo syllable with a context of two syllables*

8. RESULTS

The artificially generated pitch contours were tested within PSOLA resynthesis. The presented nn comes with very good results reading aloud newspaper texts. The intonation characteristic of our data base speaker and most of its charm is preserved. For short sentences the result is close to natural speech. Sentences that consist of more than three phrase tend to sound a little boring or monotone. This listening impression encourages the assumption of the net learning one pattern which is than modified and repeated.

The memory requirement of the presented net is less than 18k bytes. The automatic segmentation of the training data base allows an fast and completely automatic integration of new speakers and tasks. Neither making use of an human expert nor assuming properties of underlying speech or language makes this method suitable for true multilinguality.

In application the change between a small set of task dependent networks offers a computational efficient possibility for high qualitative speech synthesis. The variation of speaking styles can help to reduce the somewhat boaring attitude TTS-systems are still supposed to have.

9. REFERENCES

[1] T. Bub and J. Schwinn. Verbmobil: the evolution of a complex large speech-to-speech translation system. In *ICSLP 96, Fourth International Conference on Spoken Language Processing*, volume 4, 1996.

[2] Robert Edward Donovan. *Trainable Speech Synthesis*. PhD thesis, Cambridge University Engineering Department, 1996.

[3] Gary William Flake. Square Unit Augmented, Radially Extended, Multilayer Perceptrons. Technical report, Siemens Corporate Research, Inc., Princeton.

[4] B. Heuft, T. Portele, F. Höfer, et al. Parametric description of F_0-contours in a prosodic database. In *ICPHS '95*, volume 2, pages 378–381, Saarbrücken, 1995.

[5] George J. Klir and Bo Yuan. *Fuzzy sets and fuzzy logic: theory and applications*. Prentice-Hall, London, 1995.

[6] *SENN Version 2.2 User Manual*. Siemens Nixdorf, 1996.

[7] Gerit Sonntag, Thomas Portele, and Barbara Heuft. Prosody generation with a neural network: weighing the importance of input parameters. In *ICASSP '97*, volume 2, pages 931–934, Munich, 1997.

[8] D. S. Stall. The role of text normalization in text-to-speech synthesis. In *Informationstechnik - IT*, volume 31, pages 342–350, 1989.

[9] Christof Traber. F_0 generation with a database of natural F_0 patterns and with a neural network. In G. Bailly and C. Benoit, editors, *Talking machines: theories, models, and designs*, pages 287–304, Elsevier, North-Holland, 1992.