From Proc. ICNN'94, Orlando, Vol. 7, pp. 4457–4461

# Pitch Determination Considering Laryngealization Effects In Spoken Dialogs

H. Niemann, J. Denzler, B. Kahles, R. Kompe, A. Kiessling, E. Nöth, V. Strom

*Abstract*— **A frequent phenomenon in spoken dialogs of the information seeking type are short elliptic utterances whose mood (declarative or interrogative) can only be distinguished by intonation. The main acoustic evidence is conveyed by the fundamental frequency or $F_0$–contour.**

**Many algorithms for $F_0$ determination have been reported in the literature. A common problem are irregularities of speech known as 'laryngealizations'. This article describes an approach based on neural network techniques for the improved determination of fundamental frequency. First, an improved version of our neural network algorithm for reconstruction of the voice source signal (glottis signal) is presented. Second, the reconstructed voice source signal is used as input to another neural network distinguishing the three classes 'voiceless', 'voiced non–laryngealized', and 'voiced laryngealized'. Third, the results are used to improve an existing $F_0$ algorithm.**

**Results of this approach are presented and discussed in the context of the application in a spoken dialog system.**

## I. Introduction

Spoken dialog systems for information seeking enquiries, e.g. train or plane connections, are a current research topic. The importance and the use of intonation (or prosody, or suprasegmental information) in speech recognition and understanding has been discussed, for example, in [11] [12] [13] [14] [15]. It has been shown in [10] that in such dialogs short elliptic utterances occur frequently whose mood can only be determined by prosody. For example, an utterance of an officer or an automatic system 'the train leaves at 16.30' might be answered by the client by 'at 16.30 (!)' signalling to the officer 'ok, I got it' or by 'at 16.30 (?)' signalling to the officer 'please, confirm the time'. In both cases the wording is identical, only the intonation is different. The most important cue to intonation is the course of the fundamental frequency over time, that is, the $F_0$–contour. Many algorithms for determining the fundamental frequency have been reported [6]. A fairly reliable algorithm using dynamic programming was presented in [9] and employed in a prosodically controlled dialog system [10]. Nevertheless, irregularities of speech known as *laryngealizations* or as *creaky voice* often cause errors in $F_0$ detection which in turn may cause erroneous reactions of a dialog system. Hence, further improvement is desirable.

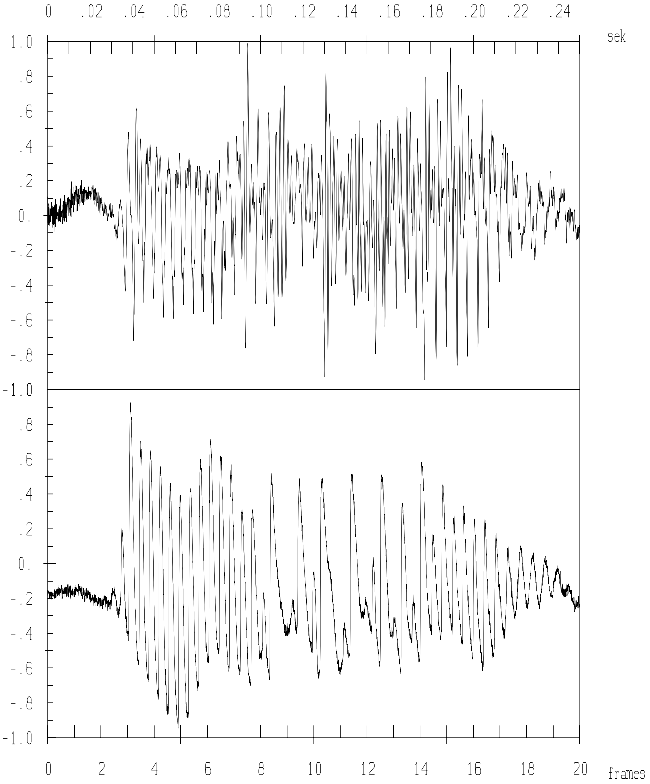


Figure 1. Top: a speech signal with laryngealization Bottom: Voice source signal, recorded with a laryngograph.

A laryngealization is usually characterized by an irregularity of the voiced excitation, manifested in the speech signal by irregular periodicity, strong variations of the amplitude, special form of the damped wave, or very long pitch periods. An example is given in Figure 1. Different types of laryngealizations are distinguished in [7] [2]. Such irregularities in the speech signal are a frequent source of errors in $F_0$ determination. However, they also occur frequently at phrase boundaries, and therefore, may be a useful cue for parsing of the utterances.

In this contribution we describe a neural network based approach to the determination of laryngealizations which in turn are used to improve $F_0$ extraction. Since many errors in the present version of the $F_0$ extraction algorithm are caused by laryngealizations, the idea is simply to detect them and to interpolate the $F_0$–contour over laryngealized portions of speech. Figure 1 indicates that laryngealizations may be detected more easily and reliably in the voice

This research was supported by the German Ministry of Research and Technology (BMFT) in the project VERBMOBIL. Only the authors are responsible for this article.

The authors are with Lehrstuhl für Mustererkennung (Informatik 5), Universität Erlangen–Nürnberg, 91058 Erlangen, Germany, except for V. Strom who is now with Institut für Kommunikationsforschung und Phonetik, Universität Bonn, 53115 Bonn, Germany

source signal than in the speech signal. Therefore, the first step is the reconstruction of the voice source signal (VSS) by means of inverse filtering. This is done by a neural network as described in Sect. II; a first version of this approach was presented in [3], and here we present an improved version of this algorithm. Once the VSS has been determined it is classified by a neural network frame by frame into one of the three classes 'voiceless', 'voiced non–laryngealized', and 'voiced laryngealized'; this is described in Sect. III. We show in Sect. IV that these results can be used to improve an existing algorithm for $F_0$ extraction. A conclusion and outlook is given in Sect. V.

## II. Inverse Filtering of the Speech Signal

### *A. Signal–To–Signal Mapping*

Since pitch period calculation and the detection of laryngealizations can be done much easier using the voice source signal (VSS) instead of the speech signal the first step is to reconstruct the VSS. Additionally, the voiced/unvoiced decision is trivial on the VSS. The approach is to map the speech signal *directly* to the VSS, i.e. to apply a signal–to–signal transformation. The input and desired output values are normalized to the range of $[0,1]$. One window (containing 78 sample values corresponding to 39 ms) of the speech signal is presented to the input layer of an artificial neural network (ANN) which gives as output one single value of the VSS. If we denote, in a multilayer–perceptron,

- the *input layer* by index $l = 0$ having $i = 0, 1, \ldots, M_0 - 1$ input nodes $n_i^{(0)}$ for signal values $x_i$,
- the *hidden layers* by indices $l = 1, 2, \ldots, M_{L-1}$ having $i = 0, 1, \ldots, M_l - 1$ hidden nodes $n_i^{(l)}$,
- the *output layer* by index $l = L$ having — in our case — *one* output node $n_0^L$,

then the relevant equations for computing an output value are:

1. for each layer and each node per layer compute the weighted sum of input values

$$y_j^{(l+1)} = \sum_{i=0}^{M_l-1} w_{ij}^{(l+1)} f_i^{(l)} - w_j^{(l+1)}$$
$$0 \le j \le M_{l+1} - 1\ , \quad l = 0, 1, 2\ , \quad (1)$$

2. compute the output of a node from a *nonlinearity* $\Theta$, in our case the sigmoid function

$$\begin{aligned} f_j^{(l+1)} &= \Theta\left[y_j^{(l+1)}\right] \\ &= \frac{1}{1 + \exp(-y_j^{(l+1)})} \end{aligned} \quad (2)$$

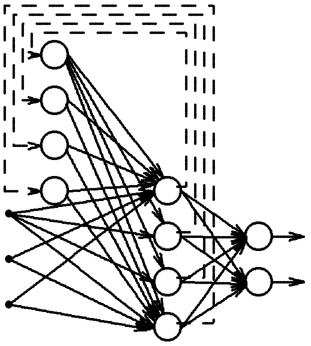
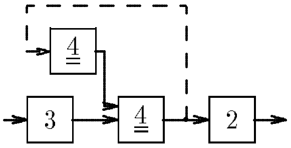


Figure 2. An example of an Elman recursive network having one hidden layer of nodes.



By shifting the speech signal point–by–point through the input layer, one gets the complete VSS reconstructed by the ANN; the reconstructed VSS is termed here ANN–VSS. Three types of ANN's are investigated, the (non–recursive) multilayer perceptron and the (recursive) Elman and Jordan networks [4] [8]; for each type different configurations were evaluated. As an example we give the structure of an Elman network in Figure 2 which has only one hidden layer of neurons. The output of every hidden neuron is fed back to a recursive node which in turn feeds forward to every hidden node. Hence, the number of hidden nodes and recursive nodes is the same in this network type. There may be additional layers of hidden nodes which in fact were used in our experiments.

We chose the sigmoid function as the activation function of all the neurons. The network is trained with the quickpropagation algorithm [5]. For training the mean–square–error criterion is used, for testing of the quality of the ANN–VSS we use coarse errors in estimation of $F_0$. If we denote the iteration step during training by $n$, the backpropagation training adjusts weights according to

$$w_{ij,n+1} = w_{ij,n} + \alpha \Delta w_{ij,n+1} + \beta \Delta w_{ij,n} \quad (3)$$

where $\alpha$ is the learning rate and $\beta$ is the momentum term . The quickpropagation algorithm replaces this by

$$\begin{aligned} w_{ij,n+1} &= w_{ij,n} + \Delta w_{ij,n} \\ \Delta w_{ij,n} &= \frac{\frac{\partial \epsilon}{\partial w_{ij,n}}}{\frac{\partial \epsilon}{\partial w_{ij,n-1}} - \frac{\partial \epsilon}{\partial w_{ij,n}}} \Delta w_{ij,n-1}\ ; \end{aligned} \quad (4)$$

hence, no choice of $\alpha$ and $\beta$ is necessary. The term $\epsilon$ in the above equation denotes the mean–square–error between input and output values. In the cases $n = 0$ and $\Delta w_{ij,n-1} = 0$ , $\frac{\partial \epsilon}{\partial w_{ij,n}} > 0$ we use equ. (3).

The speech signal is low–pass filtered with 1000 Hz; to reduce high and low frequencies the VSS is band–pass filtered from 20 Hz to 1000 Hz; and finally the signals are downsampled to 2 kHz, to reduce the amount of data. The processing simply consists of shifting the speech signal point–by–point through the input layer of the ANN, and concatenating the sequence of single output values of the ANN to the ANN–VSS. The output value of the net is interpreted as the value of the ANN–VSS, measured at the middle of the input frame. We use this input/output relationship because the other possible form of relationship, i.e. mapping one frame of the speech signal to one frame of the ANN–VSS, enlarges the net and so the training time. Furthermore, the mapping would be more complex and cause discontinuities at output frame boundaries. The

ANN–VSS is smoothed to reduce noise. We use iteratively 5 average filters the width of which depends on the average pitch period of the ANN-VSS, so that only noisy parts and no period in the ANN-VSS will be smoothed. The average period is estimated by analyzing every frame of the ANN-VSS in the frequency domain, searching for the maximum in the spectrum. Unvoiced frames are ignored for this estimation.

To train the network we use a database S in which speech and voice source signals were recorded in parallel [1]. We got a data set of 114 pairs of speech and voice source signals (recorded by a laryngograph) with sampling frequency of 16 kHz. Within the data set 3 male and 5 female speakers spoke German time of day expressions (for example, "sechzehn Uhr vier", 16.04 o'clock). The database S has a length of 140 seconds of speech. It is divided into three subsets, S1, S2, S3.

The subset S1 consists of 35 sentences from 8 speakers, 40 seconds of speech, 2757 frames, 68620 training patterns; it is used for training the various networks. The subset S2 has a non–laryngealized utterance from each of the 8 speakers. It consists of 1014 frames and is used to test (during training) the ability of a network to discriminate non–laryngealized utterances which were not in the training set. The subset S3 contains from each of the 8 speakers one utterance having at least one laryngealization. It consists of 641 frames and is used to test the ability of the network to discriminate laryngealized utterances which were not in the training set.

The mean–square–errror (MSE) is used to optimize the weights and to judge the quality of the mapping performance by an ANN. However, in this case the MSE is not an useful measure of the quality of the ANN-VSS. Some visually good signals have a greater mean square error than visually poor signals. Hence, the quality of the ANN–VSS is measured in the following way. We first calculate the pitch period of the VSS recorded by a laryngograph on a frame–by–frame basis using a modification of the algorithm given in [1] that searches for relevant maxima in the VSS. The pitch period of an ANN–VSS frame (length of a frame: 12.8 msec) is the average of the distances between all consecutive maxima in the frame. We define an *error* in one frame of the ANN–VSS if its pitch period differs from the reference (created automatically and hand-corrected) by more than 30 Hz. Thus the error is given frame–by–frame not point–by–point. This is done only for frames of voiced speech. This measurement is also close to the intuitive judgment of a person who visually analyzes the ANN–VSS.

The weights are initialized by small random numbers. The ANN is trained a fixed and relatively small number of epochs (15 epochs) with the training set S1. Then we test the ANN with the sets S1, S2, and S3 and record the number of errors. If the error becomes zero or if it increases over several (e.g. 3) iterations, the training stops and otherwise a new cycle is entered. After training the network we compute the error rate, that is the percentage of deviations in $F_0$ which are larger than 30 Hz, on the full database S.

| T | I | L | H | O | Rec | W | Error |
|---|---|---|---|---|---|---|---|
| ML | 78 | 3 | 80 | 1 | NIL | 19361 | 184 (3.5 %) |
| ML | 78 | 3 | 100 | 1 | NIL | 28201 | 159 (3.0 %) |
| JO | 78 | 2 | 100 | 1 | 10 | 19131 | 189 (3.6 %) |
| EL | 78 | 2 | 100 | 1 | 200 | 38101 | 186 (3.6 %) |

Table 1. Summary of main results.
T denotes the type of ANN (ML: multilayer perceptron, JO: Jordan recursive network, EL: Elman recursive network); I, H, and O are the number of input nodes, hidden nodes per layer, and output nodes, respectiveley; L is the number of hidden layers; Rec is the number of recursive nodes; W is the number of weights in the ANN; Error gives the absolute number of errors and the percentage.

### B. Results

As mentioned above, three types of networks were investigated. Several configurations with different number of hidden layers, different number of nodes per hidden layer, and different number of recursive nodes (for Jordan and Elman networks) were considered. Training times for larger networks were about 3 days on a workstation with about 100 MIPS. The best configurations are summarized in Table 1. This table shows that very reliable $F_0$ estimation is possible from the ANN–VSS.

An example of a reconstructed ANN–VSS is given in Figure 3. It supports the result given in Table 1 that reconstruction is very good. The irregularites in the voice source signal due to laryngealizations are clearly reconstructed, see frames 32 – 38 in the lower part of Figure 3. The results support the assumption that there are general regularities for mapping a voice source signal to a speech signal, that the inverse mapping exists, and that at least a very good approximation of this inverse mapping can be learned by an appropriate neural network.

Probably most of the remaining errors are caused by voiced/unvoiced transitions and by laryngealizations. The inverse filtering is robust to untrained speakers, different recording conditions, facilities, and vocabularies. Although a large number of network configurations was tested, it may be expected to obtain still better results with larger networks and more training data.

## III. Determination of Laryngealizations

Having reconstructed the VSS, the next step is to use it to detect laryngealizations in the speech signal. This is done by a second artificial neural network (ANN) which again is a multilayer perceptron.

The input is the ANN–VSS obtained as described in Section II above. It has a sampling frequency of 2 kHz and is normalized to the interval $[-1, +1]$. Experiments were carried out using 3, 5, and 7 frames (corresponding to 77, 128, and 179 sample values, respectively) as input to an ANN. Training of the ANN's was done with 1329 sentences

[1] This database was kindly provided by the Institute of Phonetics of the L.M. Universität, München.

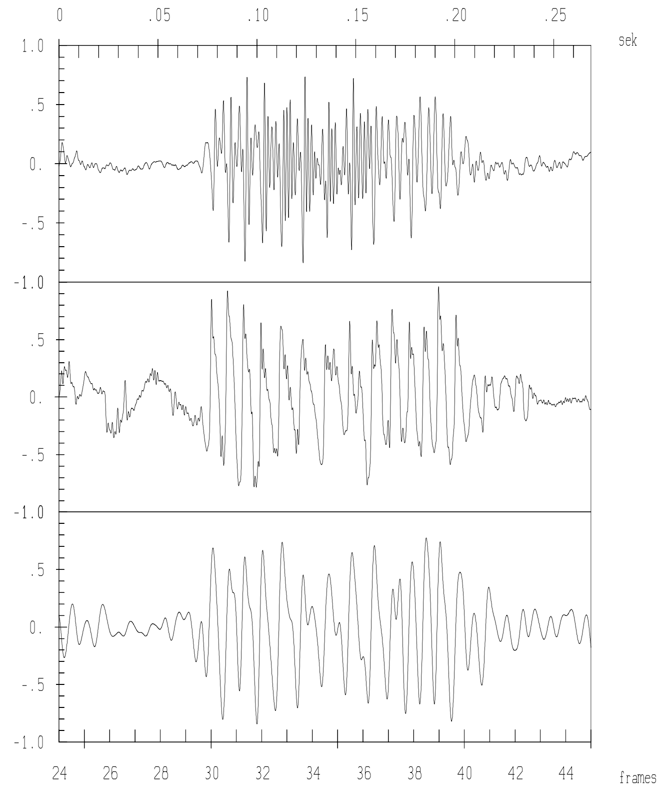


Figure 3. From top to bottom: speech signal, laryngograph VSS, ANN–VSS

(about 30 minutes) of speech spoken by 1 male and 3 female speakers, one third spontaneous and two thirds read speech. Frames of 12.8 ms duration containing laryngealizations were hand–labeled by phoneticians at L.M. University Munich. This hand–labeling is here assumed to be correct. In a first series of experiments the output of the ANN was one out of the three classes 'unvoiced', 'voiced non–laryngealized', and 'voiced laryngealized'. In a second series unvoiced frames were excluded a priori and only the two classes 'voiced non–laryngealized' and 'voiced laryngealized' were distinguished. The experiments are summarized in Figure 4.

Again, various ANN's were tried. It was found to be useful to smooth results with a median filter of width three. The best network consisted of 128 input nodes, two hidden layers each one with 80 nodes, and 3 output nodes (128–80–80–3 ANN); the results are summarized in the confusion matrix given in the upper part of Table 2. In particular, 85.6 % of laryngealized frames are classified correctly, and the false alarm rate consists of 16.1 %. This is considered to be a reasonable compromise between not missing laryngealizations (maximizing correct recognition) and minimizing false alarm rate. Using the hand labeling the unvoiced frames were excluded and a second series of experiments was made. Now the best network consisted of 128 input nodes, two hidden layers each with 100 nodes, and 2 output nodes. The results are given in the lower part of Table 2.

```
initialize weights by small random numbers, iteration step number $N = 0$, learning rate $\alpha = 0.01$, and correct classifications $p_c = 0$
    $N \leftarrow N + 1$
    train ANN on training set having an equal number of patterns from the three classes ($\approx 6000$) for a fixed number of epochs (here: 25)
    test on test set of 144 sentences and store $p_{c,N}$, that is correct classifications at iteration step number $N$
    IF    $p_{c,N} \leq p_c$
    THEN  $\alpha \leftarrow \alpha/10$
    ELSE  $p_c = p_{c,N}$
UNTIL learning rate $\alpha \leq 0.000\,01$
determine iteration step number with maximal recognition rate $p_{c,N}$ and corresponding ANN
determine with this ANN the recognition rate on the full sample of speech
```

Figure 4 Training and testing for the detection of laryngealizations.

| | classified as | | | | | |
|---|---|---|---|---|---|---|
| | UV | | VN | | VL | |
| | # | % | # | % | # | % |
| UV | 47820 | 80.2 | 6327 | 10.6 | 5456 | 9.2 |
| VN | 5320 | 6.9 | 54906 | 71.5 | 16559 | 21.6 |
| VL | 431 | 7.0 | 456 | 7.4 | 5273 | 85.6 |

| | | classified as | | | |
|---|---|---|---|---|---|
| | | VN | | VL | |
| VN | | 56929 | 74.1 | 19856 | 25.9 |
| VL | | 424 | 6.9 | 5736 | 93.1 |

Table 2. Summary of main results for detection of laryngealizations. The abbreviations are UV for 'unvoiced', VN for 'voiced non–laryngealized', and VL for 'voiced laryngealized'. The upper part of the table shows results on three classes, the lower part on two classes.

The recognition rate could be increased to 93.1 %, but the false alarm rate increased to 25.9 %.

In Sect. II we showed an approach to inverse filtering using a neural network. The "classical" approach to inverse filtering is a linear filter and, of course, the question arises, whether anything can be gained from using an ANN instead of a linear filter. Hence, in another experiment we computed the voice source signal from a linear inverse filter and used this VSS to train a 128–80–80–3 ANN, which was the best network type for the ANN–VSS. The result is that 75.9 % of laryngealized frames are correctly classified (instead of 85.6 % for the neural network inverse filter), that the mean recognition rate on all frames drops to 63.1 % (from 76.5 %), and that the false alarm rate is almost the same for both approaches (about 16 %). This result clearly demonstrates the superiority of ANN inverse filtering to linear inverse filtering for the detection of laryngealizations.

| | NO | REF | ANN |
|---|---|---|---|
| UV/VO error in % | 3.9 | 3.9 | 3.8 |
| VO/UN error in % | 6.9 | 7.3 | 8.3 |
| coarse $F_0$ error % | 10.4 | 6.5 | 7.2 |
| av. error in Hz | 17 | 11 | 12 |

Table 3 Improved computation of the fundamental frequency by exclusion of laryngealized sections of speech. The abbreviations are: NO for no consideration of laryngealized sections, REF for using the hand labeled laryngealizations, and ANN for using the automatically (ANN) detected laryngealizations.

## IV. Improved Pitch Determination

As mentioned in the introduction (Sect. I), the final goal is a reliable estimate of the fundamental frequency $F_0$. An algorithm for $F_0$ computation based on dynamic programming has been described elsewhere [9]. It first computes several candidate values for the fundamental frequency using two independent algorithms and then computes an optimal $F_0$ contour by dynamic programming. A problem are wrong candidate values which often are caused by irregularities of speech, that is laryngealized frames. Therefore, the idea is to detect laryngealizations, as shown above, and then to exclude those frames from the DP–based computation of the fundamental frequency by labeling them as "unvoiced". After computation of the $F_0$ contour this contour is interpolated linearly in laryngealized frames; if the laryngealized section is at the beginning or end of a voiced section, the $F_0$ value of the last voiced frame is extrapolated by a constant.

The results for the database of 1329 sentences are summarized in Table 3. The table shows that the coarse error rate, that is fundamental frequency errors of more than 30 Hz, is reduced significantly (from 10.4 % to 7.2 %). Furthermore, there is only a small degradation of performance when using laryngealizations automatically detected by the ANN instead of the hand labeled laryngealizations (from 6.5 % coarse errors to 7.2 %). This result indicates the overall success of reconstructing the voice source signal and detecting laryngealizations by means of neural networks.

## V. Conclusions and Outlook

In this paper we showed that the voice source signal can be reconstructed reliably from the speech signal using an artificial neural network. With a network of type (78–100–100–100–1) we achieved a reconstruction giving 3.0 % of coarse errors in fundamental frequency computation. The reconstructed VSS was used to detect laryngealized sections of speech using another neural network. With a network of type (128–80–80–3) 85.6 % of laryngealizations are classified correctly at a false alarm rate of 16.1 %. In addition it was shown that a voice source signal computed from a *linear* inverse filter performs worse. Finally, it was demonstrated that the performance of a dynamic programming based algorithm for the computation of $F_0$ contours was improved from 10.4 % coarse errors to 7.2 % coarse errors.

We expect that these results will make prosodic control of a spoken dialog system more reliable and in turn improve its user acceptance. However, verification of this expecation will be the subject of future work.

## References

[1] P. Alku. An automatic method to estimate the time based parameters of the glottal pulseform. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages II, 29–32, San Francisco CA, 1992.

[2] A. Batliner, S. Burger, B. Johne, A. Kiessling. M"USLI: A Classification Scheme For Laryngealizations. In *Proc. ESCA Workshop on Prosody*, pages 176–179, Lund, Sweden, 1993

[3] J. Denzler, R. Kompe, A. Kiessling, H. Niemann, E. Nöth. Going back to the source: Inverse filtering of the speech signal with ANN's. In *Proc. European Conf. on Speech Communication and Technology*, pages 111–114, Berlin, Germany, 1993.

[4] J.L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

[5] S. Fahlman. An empirical study of learning speeds in backpropagation networks. Technical Report CMU-Cs-88-162, Carnegie Mellon University, Pittsburgh PA, 1988.

[6] W. Hess. *Algorithms and Devices for Pitch Determination of Speech Signals.* Springer, Berlin, 1983.

[7] D. Huber. *Aspects of the Communicative Function of Voice in Text Intonation.* PhD Thesis, Chalmers University, Göteborg/Lund, Sweden, 1988

[8] M. Jordan. Attractor dynamics and parallelism in a connectionist sequential machine. In L. Erlbaum, editor, *Proc. 1986 Cognitive Science Conference*, pages 531–546, 1986.

[9] A. Kiessling, R. Kompe, H. Niemann, E. Nöth, A. Batliner. DP–based determination of f0 contour from speech signals. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pages II–17 – II–20, San Francisco, CA, 1992.

[10] R. Kompe, A. Kiessling, T. Kuhn, M. Mast, H. Niemann, E. Nöth, K. Ott, A. Batliner. Prosody takes over: A prosodically guided dialog system. pages 2003–2006, Berlin, Germany, 1993.

[11] W.A. Lea. *Intonational Cues to the Constituent Structure and Phonemics of Spoken English.* PhD thesis, Indiana University, 1972.

[12] E. Nöth. *Prosodische Information in der Sprachverarbeitung, Berechnung und Anwendung.* PhD thesis, Technische Fakultät, Universität Erlangen-Nürnberg, Erlangen, 1990.

[13] J. Vaissiere. The use of prosodic parameters in automatic speech recognition. In H. Niemann, M. Lang, G. Sagerer, editors, *Recent Advances in Speech Understanding and Dialog Systems*, volume 46 of *NATO ASI Series F*, pages 71–100. Springer, Berlin, 1988.

[14] A. Waibel. *Prosody and speech recognition.* PhD thesis, Carnegie-Mellon Univ. Pittsburgh, USA, 1986.

[15] M.Q. Wang, J. Hirschberg. Automatic classification of intonational phrase boundaries. *Computer Speech & Language*, 6:175–196, 1992.