BMFT
gefördert vom
Bundesministerium
für Forschung
und Technologie

Verbmobil
Verbundvorhaben

GOING BACK TO THE SOURCE: INVERSE FILTERING OF THE SPEECH SIGNAL WITH ANNs

J. Denzler, R. Kompe
A. Kießling, H. Niemann
E. Nöth

F.–A.–Universität Erlangen–Nürnberg

Vm
Report 32
Oktober 1994

Oktober 1994

J. Denzler, R. Kompe
A. Kießling, H. Niemann
E. Nöth

Lehrstuhl für Mustererkennung (Inf. 5)
Friedrich–Alexander–Universität Erlangen–Nürnberg
Martensstr. 3
D–91058 Erlangen

Tel.: (09131) 85 - 7890
e-mail: {denzler,kompe}@informatik.uni-erlangen.de

Gehört zum Antragsabschnitt: 3.11, 3.12

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 H/0 und 01 IV 102 C 6 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

From Proc. EUROSPEECH 1993, Berlin, Vol. 1, pp. 111–114

GOING BACK TO THE SOURCE: INVERSE FILTERING OF THE SPEECH SIGNAL WITH ANNs

J. Denzler, R. Kompe, A. Kießling, H. Niemann, E. Nöth
Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung (Informatik 5), Martensstr. 3, 91058 Erlangen, FRG
phone: +49/9131/857874, fax: +49/9131/303811, e-mail: denzler@informatik.uni-erlangen.de

Abstract

In this paper we present a new method transforming speech signals to voice source signals (VSS) using artificial neural networks (ANN). We will point out that the ANN mapping of speech signals into source signals is quite accurate, and most of the irregularities in the speech signal will lead to an irregularity in the source signal, produced by the ANN (ANN-VSS). We will show that the mapping of the ANN is robust with respect to untrained speakers, different recording conditions and facilities, and different vocabularies. We will also present preliminary results which show that from the ANN source signal pitch periods can be determined accurately.

Keywords: ANN, inverse filtering.

1 Introduction

State of the art pitch period detection is mostly done using the speech signal or some frequency representation of it. Within voiced speech one wants to calculate the frequency of the opening and closing of the vocal folds. Furthermore, there is an increasing interest in the detection of laryngealizations, which are irregular but voiced excitations of the vocal folds. Two different characterization schemes for the appearance of laryngealizations in the speech signal with a fine subcategorization of laryngealizations can be found in Huber [8] and Batliner [2]. These irregularities are often the reason for incorrect pitch period detection using the speech signal. Furthermore, Huber [8] and Kießling [10] have shown that laryngealizations often occur at linguistic boundaries so that the detection of laryngealizations can be used for parsing analysis of the speech signal.

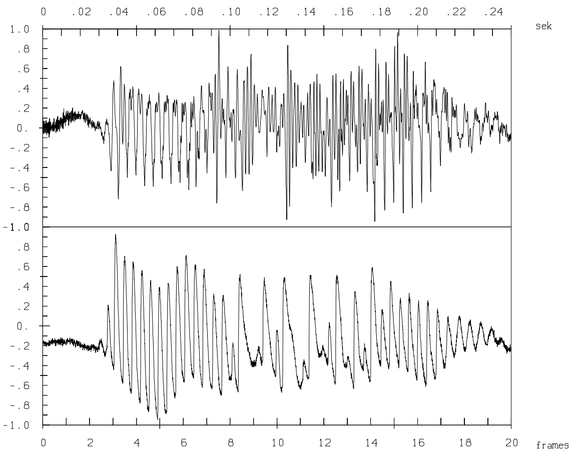

Figure 1: Top: Speech signal with laryngealization. Bottom: Voice source signal, recorded with a laryngograph.

Both pitch period calculation and the detection of laryngealizations can be done much easier using the voice source signal instead of the speech signal (see Figure 1). Additionally, the voice/unvoiced decision is trivial on the voice source signal. The voice source signal can be measured using pitch detection instruments like the laryngograph (Heß, [7]). Usually the laryngograph–voice source signal is not available to a speech recognition system. Therefore another approach is the transformation of the speech signal into the VSS by the method of inverse filtering. In this paper we present a new method for this transformation using ANNs trained on a set of speech signals for which a laryngograph signal is available.

2 Inverse Filtering with ANNs

It has been shown that ANNs can be applied to tasks like classification, signal processing or simple mapping of one data set to another. Lapedes [11] has shown that ANNs can be used for nonlinear signal processing. Most work in the field of speech recognition with ANN, for example in phoneme recognition [6], [3], pitch detection and voice/unvoiced decision (for further references see [4]), concerns feature-to-feature or feature-to-symbol transformation.

In our approach we map the speech signal directly to the VSS, i.e. signal-to-signal transformation. Therefore we do not do any coding of the signals, except a normalization of the input and output values to the range of $[-1, 1]$. This is totally different from some other work, where features were used as input values to the ANN extracted from the speech signal. We present one frame of the speech signal to the input layer of the ANN and then get one single output value at the output layer. This value is interpreted as one signal point of the VSS. By shifting the speech signal point-by-point through the input layer, we get the complete VSS by concatenating the single output values to one signal. The width of the input frame, i.e. the number of input values, and the relationship between the input frame and the output value will be described in section 4.

Like Lapedes [11] and others we use a multilayer perceptron with up to three hidden layers. Each layer n is fully connected with the layer $n + 1$. We choose the sigmoid function as the activation function of all the neurons. The network is trained with the quickpropagation algorithm [5]. We could not obtain any positive results using the backpropagation algorithm. No recurrent links were used.

3 The speech data

To train the network we needed a database, in which speech and voice source signals were recorded in parallel (database L)[1]. We got a data set of 114 pairs of speech and voice source signals (sampling frequency 16 kHz). Within the data set 8 speakers, 3 male 5 female, spoke German time of day expressions (for example "zweiundzwanzig Uhr neun"). To get a lot of laryngealized signals, two of the 8 speakers (1 male, 1 female) tried to produce irregularities during the recording. One example is shown in Figure 2. In totaly the data set has a length of 267 seconds of spoken German. We extracted from this set two subsets, which were used during the training. First we took one training set consisting of 10 sentences (10 seconds speech), from 6 of the 8 speakers (data set L1). The two speakers not used for training were used to test the network performance after the training (data set LT). 6 out of the 10 sentences contained laryngealizations. To reduce the computation time (some of our experiments need two or more weeks on a DEC-Station 5000/200), we used only this small training set. However, these 10 seconds conform to more than 20000 training patterns. Secondly we use a small test set (data base L2) of 18 sentences from the 8 speakers, containing both laryngealizied and normal spoken utterances, for testing the network performance during the training. The rest of data set L was only used after training, to make a concluding test of the network. In the total data set L, the pitch ranges from 33 Hz to 380 Hz, with a mean pitch of 152 Hz and a standard deviation of $\sigma = 62$. The training set L1 has a range from 52 Hz to 353 Hz (mean: 178 Hz, $\sigma = 64$).

To test the configured ANN with a larger database we take the so called SPONTAN-data set (S) (sampling frequency 10 kHz, 4 speakers, different recording conditions from data set L, see [10]), a data set which contains spontaneously spoken German sentences. For this speech data we did not have any voice source signal, but frame-wise hand–corrected pitch values. Additionally laryngealized frames were marked by experienced phoneticians.

4 The System

Our system is divided into three parts:

Preprocessing
To reduce high frequencies and low frequencies (in the VSS high energy low frequencies were caused by larynx movements) the VSS was band–pass filtered from 20 Hz to 1000 Hz, the speech signal low-pass filtered with 1000 Hz. Then we sample the signals down to 2 kHz, to reduce the amount of data.

Processing
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 VSS, produced by the ANN. We use the following relationship between the input frame and the output value: the output value of the net was interpreted as the value of the VSS, measured at the middle of the input frame. We used 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 VSS, enlarges the net and so the training time. Furthermore, the mapping would be more complex and one would get unsteadiness at output frame bounderies.

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

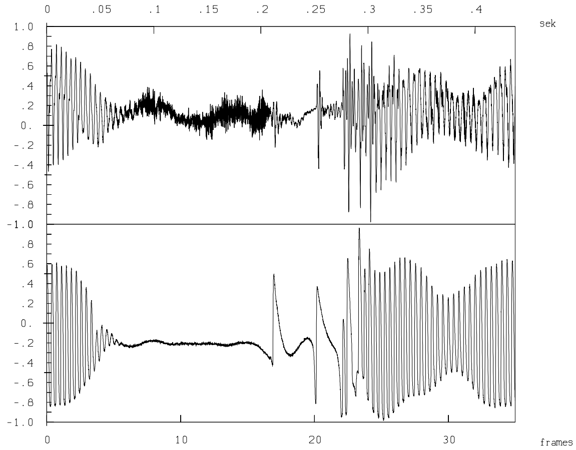

Figure 2: Example speech and voice source signal out of the data set L.

Postprocessing
To reduce noise in the VSS produced by the ANN we have to smooth it. We used iteratively 5 average filters whose width depended 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.

5 The Error Measure

In most applications of multi-layer perceptrons, the error criterion is the mean square error (MSE). The MSE is used to optimize the weights and to judge the quality of the mapping performance by the ANN. In our case the MSE is not an exact measure of the quality of the ANN-VSS. Some visually good signals have a greater mean square error than visually poor signals. Thus, the MSE is still used to optimize the weights. However, the quality of the ANN–VSS is measured in the following way.

We first calculate the pitch period of the reference VSS on a frame-by-frame basis. For that purpose, we modified an algorithm developed by Alku [1], to search for relevant maxima in the VSS. These maxima can only be found accurately if the signal is not noisy, and in a periodic form. With this pitch synchronous algorithm we calculate the pitch period of the ANN-VSS frame-by-frame. 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 a correct ANN–VSS within a frame as one whose pitch period differs from the reference (created automatically and hand-corrected) by less than 30 Hz. Thus the error is given frame–by–frame not point–by–point. This is done only for frames of voiced speech. As we will point out in section 6, this measurement is close to the intiutive judgment of a person who visually analyzes the ANN–VSS.

6 Experiments and Results

As stated earlier we used multilayer perceptrons with the quickpropagation learning rule. In our experiments we always used the sigmoid activation function. We varied the number of hidden layers, the number of input and hidden nodes. Here we will present only the most important parts of the results (further detail can be found in [4]).

First we will describe the iterative training procedure:

1. We train the ANN a fixed and relatively small number of epochs (15 epochs) with the training set L1.
2. Then we test ANN configured in this way with the training set L1 and the test set L2. We need this, to calculate the above defined error measure.
3. We compare the error rate with the error rate of the last iteration.
4. If the error grows for more than three iterations on $L1 \cup L2$ we stop the training; else we go to step 1.

When the ANN is trained all 114 utterances are taken to test the ANN (data set L). We count again the number of frames with an error. This number will be shown in the following as the result of the ANN.

In Table 1 the error rate is shown for the best net. We got this result after 225 epochs. The net has three hidden layers with 120 nodes in each layer, 78 input nodes (i.e.

Data set	Error rate in percent
L	3.5
LT	10.2

Table 1: Results of the best net on the complete data set L, and on the untrained speakers LT.

39 msec) and one output node. To train one epoch, i.e. 20000 training patterns, we need 2000 seconds on a DEC 5000/200.

In Figure 3 an ANN–VSS is shown, from a speech signal in the training set L1. One can see that the ANN has even learned to build the irregularities in the VSS at the laryngealized frames 32–38. In Figure 4 a speech signal from the data set LT is shown. This is one of the worst cases for data set LT. Nevertheless the ANN transforms the "normally" voiced frames 35–45 correctly to the VSS. Only the laryngealized parts (frames 30–36 and 45–53) are not well transformed. Further experiments will have to show whether the detection of laryngealizations can be

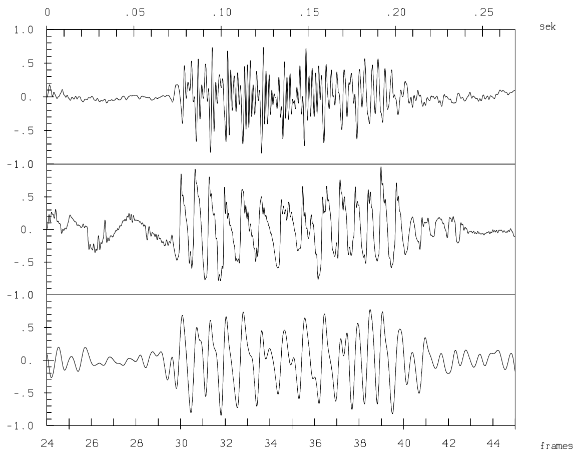

Figure 3: Speech signal, voice source signal and ANN–VSS of the training set L1.

done better in the transformed signal rather than in the original speech signal.

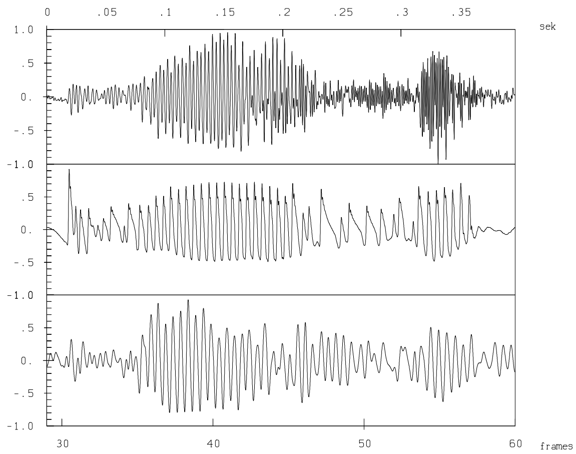

Figure 4: Speech signal, voice source signal and ANN–VSS of an untrained speaker (data set LT)

Additionally we tested our best ANN with the data set S. We do that because we can test a larger set of utterances. One can see how the ANN transforms speech signals, recorded under different conditions and facilities and spontaneous spoken from untrained speakers. Additionally this test set contains a different vocabulary than the training data set. Since we have a reference pitch period for all the utterances, we can test the ability to determine pitch correctly from the ANN–VSS.

For all four speakers of the data set we compare the number of incorrect frames with the number of incorrect frames of a pitch determination, based on the spectrum of the speech signal with the algorithm described in [9]. The pitch period determined from the ANN–VSS is on the average up to 10 % better than the pitch determined from the speech signal. This shows that the ANN–VSS can be used for pitch determination.

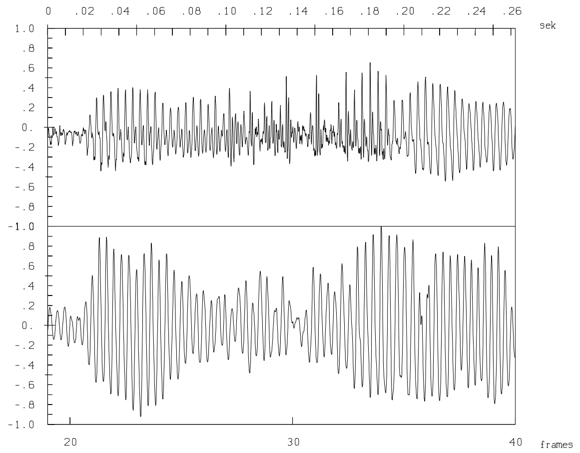

Figure 5: Speech signal and ANN–VSS of a speaker in data set S (no reference VSS available).

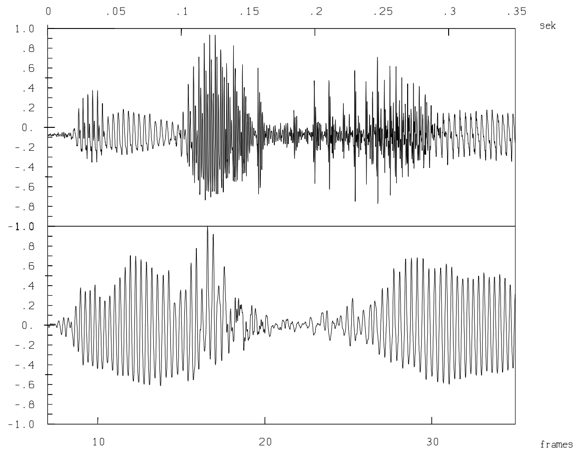

Figure 6: Speech signal and ANN–VSS of a speaker in data set S (no reference VSS available).

In Figure 5 one of the speech signals and the ANN–VSS of data set S is shown. One can see that during regular, non-laryngealized speech the ANN–VSS is quite accurate. During the laryngealization the ANN interpolates the VSS signal, i.e. no irregularities in the ANN–VSS are produced. In Figure 6 the other type of transformation at laryngealizations is shown. As in Figure 5 during regular speech the ANN–VSS is accurate. Within the frames 18–27 the ANN produces a nearly constant low energy signal. This can be useful for detecting laryngealizations.

7 Discussion

We have shown that an ANN can be trained which transforms speech signals to voice source signals quite accurately. Probably most of the errors are caused by voiced/unvoiced transitions and by laryngealizations. This inverse filtering is robust to untrained speakers, different recording conditions, facilities, and vocabularies. Plots of the ANN-VSS show that laryngealizations may also be detected in the ANN–VSS. We will investigate this in future work. Further work will also be done using larger networks and larger training sets.

Acknowledgements

This work was supported by the German Ministry for Research and Technology (*BMFT*) in the joint research project ASL/VERBMOBIL. Only the authors are responsible for the contents of this paper. We wish to thank the *Bayerisches Forschungszentrum für Wissensbasierte Systeme* (*FORWISS*), whose Neural Network simulator (FAST) was used for the experiments reported in this paper. In particular, M. Arras kindly provided support concerning software and interfaces.

References

[1] P. Alku. An automatic method to estimate the time based parameters of the glottal pulseform. In *International Conference on Acoustics Speech and Signal Processing*, number II, pages 29–32, 1992.

[2] A. Batliner, S. Burger, B. Johne, and A. Kießling. *MÜSLI: A Classification Scheme For Laryngealizations.* ESCA Workshop on prosody, Lund (Schweden), Sept. 1993.

[3] Y. Bengio, R. De Mori, G. Flammia, and R. Kompe. Phonetically Motivated Acoustic Parameters for Continuous Speech Recognition using Artificial Neural Networks. *Speech Communication*, 11(2–3):261–271, 1992.

[4] J. Denzler. Transformation von Sprachsignalen in Laryngosignale mittels künstlicher neuronaler Netze. Master's thesis, Lehrstuhl für Informatik 5 (Mustererkennung), Universität Erlangen-Nürnberg, 1992.

[5] S. Fahlman. An empirical study of learning speed in backpropagation networks. Technical Report CMU-CS-88-162, Carnegie Mellon University, September 1988.

[6] N. Hataoka and A.H. Waibel. Evaluation of Speaker-Independent Phoneme Recognition on TIMIT Database Using TDNNs. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 105–108. 1991.

[7] W. Hess. *Pitch Determination of Speech Signals.* Springer, 1983.

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

[9] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. DP-Based Determination of *F0* contours from speech signals. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages II–17–II–20, San Francisco, 1992.

[10] A. Kießling, R. Kompe, E. Nöth, and A. Batliner. Irregularitäten im Sprachsignal — störend oder informativ? In R. Hoffmann, editor, *Elektronische Signalverarbeitung*, volume 8 of *Studientexte zur Sprachkommunikation*, pages 104–108. TU Dresden, 1991.

[11] A. Lapedes and R. Farber. Nonlinear signal processing using neural networks: Prediction and system modelling. Technischer Bericht LA-UR–87-2665, Los Alamos Laboratory, 1987.