PULSE-INDUCED MASKING OF SONAGRAMS FOR THE ENHANCEMENT OF SPEECH

Markus Volkmer and *Boris Banke*
Department of Distributed Systems, Technical University Hamburg-Harburg,
Schwarzenbergstrasse 95, D-21073 Hamburg, Germany
Phone: +49 (0)40 42878 3357, Fax: +49 (0)40 42878 2798
e-mail: {markus.volkmer,banke}@tu-harburg.de

ABSTRACT

Speech-enhancement on the basis of sonagrams and an artificial neural network is investigated. We utilise a biologically inspired model – a Pulse Coupled Neural Network – and its temporal segmentation property. Pulses of linked neurons and their temporal as well as their spacial relation regarding the receptive field induce masks, providing characteristic elements for a reduced signal-representation. Properties of the model and its integration into an enhancement process are presented. Experiments on speech and non-speech signals are evaluated and discussed with regard to observed signal processing capabilities.

1 INTRODUCTION

The enhancement of speech (or audio signals in general) is an important and necessary preprocessing step in many applications, for which also artificial neural networks (ANNs) are used [1]. It is primarily motivated by the need to improve the perceived quality of speech in the presence of noise, or to generally increase speech intelligibility. Segmented signals, for example, provide crucial information for feature- and phoneme-extraction purposes. These again serve as input to subsequent recognition or classification algorithms. An unprocessed and noisy input at an early stage often propagates poor results.

Spectrograms (or sonagrams in the context of audio signals) are applied in fields dealing with time-varying signals, including speech analysis and auditory display research. Gained from a windowed fourier transform and used as a representation of speech, they contain many details that are not relevant to encode the linguistic information.

This paper describes an ANN-approach to the enhancement of speech on the basis of sonagrams. It uses specific properties of the structure and the dynamics of a Pulse Coupled Neural Network (PCNN) model [2] to provide elements for a reduced signal-representation. After a brief review of the PCNN model underlying the approach and its origin (section 2) follows the presentation of the sonagram-based enhancement (section 3). Experimental results (section 4) illustrate the properties of the approach. We conclude with a discussion and implications on further research (section 5).

2 A PULSE COUPLED NEURAL NETWORK

Pulsed Neural Networks subsume ANN models, which take into account the neuro-biological finding that biological neural networks use pulses, and especially their timing, to code information [3]. A particular category – the PCNN – models the network as a dynamic system with no explicit learning paradigm involved, yet adapting internal parameters. Effects of collective excitation of the modeled neurons produce signal processing capabilities. Its mathematical abstraction has similarities to Cellular Automata and Coupled Map Lattices [4]. Originating from a system of differential equations [5], and being derived from a model used to describe the neural activity in visual cortex of the cat [6], the PCNN represents a discrete simplification of such a regular neural structure. Having emerged in the visual domain, it has primarily found applications in the field of image processing [7].

In our research, we address the audio signal processing capabilities of the PCNN with regard to the temporal segmentation of a sonagram, leading to pulse-induced masking.

2.1 The applied neuron model

We first introduce the neuron model (Figure 1) adopting the notation used in the IEEE Special Issue [2]. A neuron ij consists of a feeding compartment F_{ij}, a linking compartment L_{ij}, a (dynamic) threshold Θ_{ij} and a pulse generator described below. Each neuron receives a direct stimulus S_{ij}

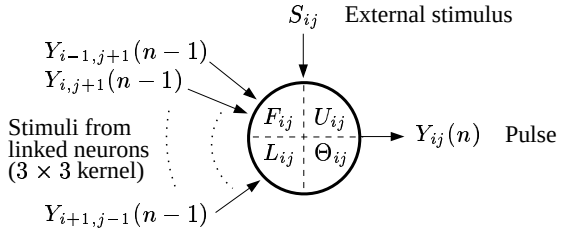

Figure 1: *Schematic diagram of a pulse coupled neuron. Only stimuli and output pulse at iteration n are shown. Internal dynamics are described in section 2.2 below.*

from the receptive field and additional stimuli via links to neighbouring neurons. In accumulating the received stimuli, the internal activation U_{ij} reaches the dynamic threshold

and the neuron emits a pulse. This relates to the notion of a *Threshold and Fire* pulse generator (cf. [3] and [8]).

Neurons are arranged in a single layer as a regular two dimensional lattice. Lateral connections are spatially defined via a kernel function, mapping a (weighted) local neighbourhood structure. The resulting network structure leads to dynamics that are described in the following.

2.2 The dynamics of the network as an iterative system

To briefly describe the underlying discrete time model of the applied PCNN, let n indicate discrete time steps (iterations). V_L, V_F and V_Θ denote three constant potentials. $e^{-\alpha_F}$, $e^{-\alpha_L}$ and $e^{-\alpha_\Theta}$ are weights with regard to the influence of the previous iteration. Matrices m_{ijkl} and w_{ijkl} are Gaussian kernels, defining the local neighbourhood (interconnection structure). Note that these parameters are identical to all neurons.

With S_{ij} denoting the external stimulus to neuron ij in the receptive field, the total stimulus to a neuron is given by

$$F_{ij}(n) = S_{ij} + e^{-\alpha_F} \cdot F_{ij}(n-1) + V_F \sum_{kl} m_{ijkl}\, Y_{kl}(n-1)\,. \quad (1)$$

$$L_{ij}(n) = e^{-\alpha_L} \cdot L_{ij}(n-1) + V_L \sum_{kl} w_{ijkl}\, Y_{kl}(n-1) \quad (2)$$

denotes the linking compartment. The excitation potential

$$U_{ij}(n) = F_{ij}(n) \cdot (1 - \beta L_{ij}(n)) \quad (3)$$

(internal activity) is proportional to the product of F and L, weighted by the linking coefficient β. Pulses Y are generated according to the excitation U and the dynamic threshold Θ:

$$Y_{ij}(n) = \begin{cases} 1\,, & U_{ij}(n) > \Theta_{ij}(n) \\ 0\,, & \text{otherwise} \end{cases} \quad (4)$$

with

$$\Theta_{ij}(n) = e^{-\alpha_\Theta} \cdot \Theta_{ij}(n-1) + V_\Theta\, Y_{ij}(n)\,. \quad (5)$$

The discrete time model of the PCNN is computed in iterating equations 1-5.

2.3 Some properties of the neural network

The special structure and the dynamics of the PCNN lead to signal processing properties:

- Inherent dynamic thresholding (eq. 5)
- Filtering through local linking-structure (eqs. 1 and 2)
- Temporal segmentation of a 2D-signal through synchronisation of firing-times and feature binding

Note that temporal segmentation [6] is not to be confused with a phonetic (1D-)segmentation. Other parameter dependant properties comprise shifting from segmentation to gradient detection or smoothing, up to the extraction of local pattern information (texture).

For a more detailed description of PCNNs and their properties, the reader is referred to a special issue of the *IEEE Transactions on Neural Networks* [2].

3 THE MASKED SONAGRAM REPRESENTATION

The human auditory system decomposes incoming signals into their constituent frequencies via the space-frequency transfer function of the cochlea [9]. Moreover, by technical means, a short time fourier transform (STFT) is used to produce a time-frequency domain representation of the signal: the sonagram. Our approach uses sonagrams the temporal segmentation property of the PCNN to achieve a separation of spectro-temporal regions. These regions are again used to create a reduced signal-representation for enhancement.

3.1 Network stimulus from preprocessing

The original time domain signal is processed, applying a STFT with a Hamming-window for the reduction of artifacts. Window-type and -size as well as sampling rate and signal length determine the size of the sonagram; and consequently the size needed for the receptive field of the network.

3.1.1 A receptive field for a sonagram

Mapping the discrete time-frequency representation bijectively to a receptive field of equal dimension that receives the localised amplitudes as stimuli, we define input and size of the PCNN. In the resulting two-dimensional structure of the network, each neuron is associated with a unique location in time-frequency space. The interconnection structure (or kernel) determines a local neighbourhood in that space.

3.2 Pulses masking time-frequency regions

Simulating the so formed network in computing the equations 1-5 (section 2.2) in discrete time steps, each neuron may emit a pulse, depending on the accumulated stimuli and the current threshold. This leads to a discrete time-series of pulses. Neurons with highest internal activity U_{ij} fire first, followed by neurons with lesser internal activity.

Pulses $Y(n)$ of the whole network can be interpreted as masks to be applied to the original sonagram representation S of the given signal, with each mask coding a particular trait in the receptive field. They represent homogeneous regions (segments) of similar energy within the sonagram. This *pulse-induced masking* can be formalised as an element-wise product of the pulse-induced mask $Y(n)$ at a given iteration n and the original external network stimulus (the sonagram) S :

$$Y_{ij}(n) \cdot S_{ij} =: \hat{S}_{ij}(n) \quad ;\, \forall i, j\,. \quad (6)$$

It results in the *masked sonagram representation* $\hat{S}(n)$. An OR-operation on masks $Y(\tilde{n})$ can be used to compose an enhanced signal

$$S_{ij} \approx \tilde{S}_{ij} := \left(\bigvee_{\tilde{n} \in 1,\cdots,n} Y(\tilde{n}) \right)_{ij} \cdot\, S_{ij} \quad ;\, \forall i, j\,, \quad (7)$$

where particular $\tilde{n}$ could be selected according to a segment-quality measure. Note that successive masks may contain redundant information. Figure 2 sketches the complete process. Properties of the temporal segmentation are completely defined by the external network stimulus S and the internal parameters of the PCNN model. In particular kernel-size and

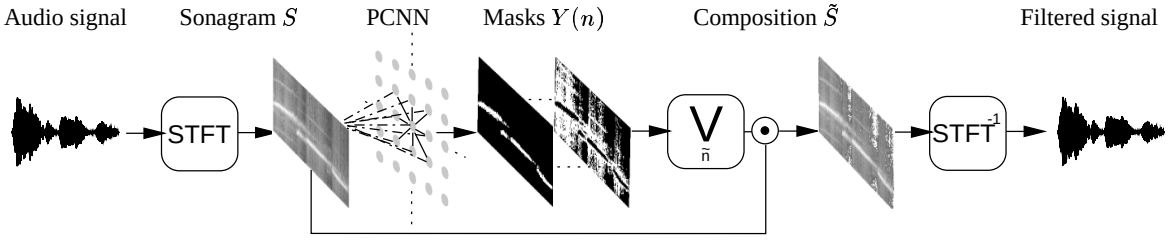

Figure 2: *The enhancement process. Local neuron interconnection structure (magnified) and composition are sketched.*

-shape, the weights and the potentials influence the behaviour and output of the network. Such effects are described more closely in [2] and [3].

4 EXPERIMENTAL RESULTS

The original audio data was down-sampled to 12 kHz (Mono) and 16-bit resolution. A Hamming-window with an overlap of 50 percent (64 samples) was used for a 128-tap FFT, resulting in a sonagram with a frequency range of 6 kHz and 64 samples per spectrum. No prior de-noising or filtering was performed. Sonagrams are displayed in grey-scales, indicating the energy of the signal proportional to brightness on a logarithmic scale. Black regions indicate complete masking. The parameters used in the PCNN were the following: $V_L = 0.15$, $V_F = 0.15$, $V_\Theta = 13.0$, $\alpha_F = 0.1$, $\alpha_L = 1.0$, $\alpha_\Theta = 0.2$ and $\beta = 0.1$. The experimental results were generated on a 166 MHz Pentium 1 with the PCNN implemented in C. A typical run of 19 iterations, for a 64 $\times$ 161 input (sonagram) with a 5 $\times$ 5 neighbourhood, took 1.5 seconds.

We display the original time domain signal (and a provided phonetic segmentation in case of speech) along with the generated sonagram S, masked sonagrams $\hat{S}$ and a composition $\tilde{S}$ from the given masks.

4.1 A Speech signal

The signal displayed on top of Figure 3 (10262 samples) is the Dutch utterance *'hartstikke leuk' (great fun)*, taken from the IFA Spoken Language Corpus [10], which is an open-source database of hand-segmented Dutch speech.

$\hat{S}(7)$ very well masks formants of the utterance. It yields no notable difference to the original signal regarding the hearing impression. Obviously, the major features are preserved by only a coarse subset of the energy distribution. Finer details are separated and represented in other masks. The voice bar remains and fricative /s/ is visible in the upper frequency range. Lateral /l/ and diphthong /ø/ are intact, though the original sonagram contains much more information. It appears that redundancy in speech admits to reduce the representation significantly, merely using spectro-temporal components and their contribution for overall intelligibility. $\hat{S}(8)$ keeps the low-energy envelope to the basic sound, but contains much of the redundant parts present in S. In $\hat{S}(9)$ another significant portion of the signal is masked, containing low-energy noise across almost all frequencies. Mask $Y(11)$, leading to $\hat{S}(11)$ (not shown here), contains

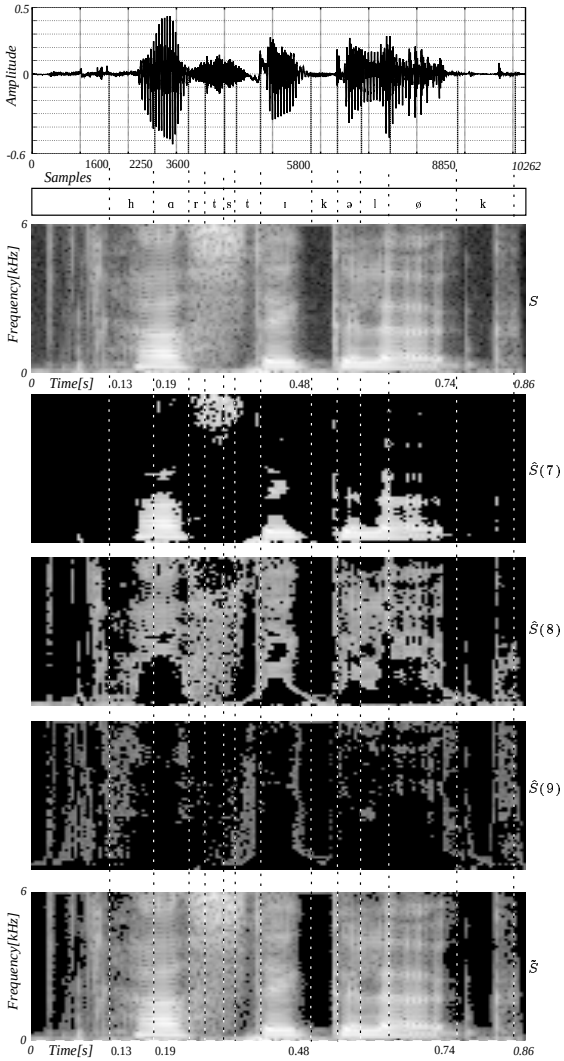

Figure 3: *Processing result (speech).*

almost exactly the sibilants. A composition $\tilde{S}$ of all three masks almost resembles the original signal, though the silence phase of the plosives remain masked.

4.2 A non-speech signal

To give further and comparative illustration, we also present results on a non-speech signal of 10396 samples (top of Figure 4). It represents a part of a song of a *banded wren (Thryothorus pleurostictus)*. It is available from the Bioa-

coustics Research Program of the Cornell Lab of Ornithology and their studies on banded wren vocal communication [11]. In contrast to a human speech signal, it consists of much clearer, yet not too simple frequencies, which help to visualise the algorithms properties. Here, we marked characteristic points in time where a shift or a change in frequency occurs.

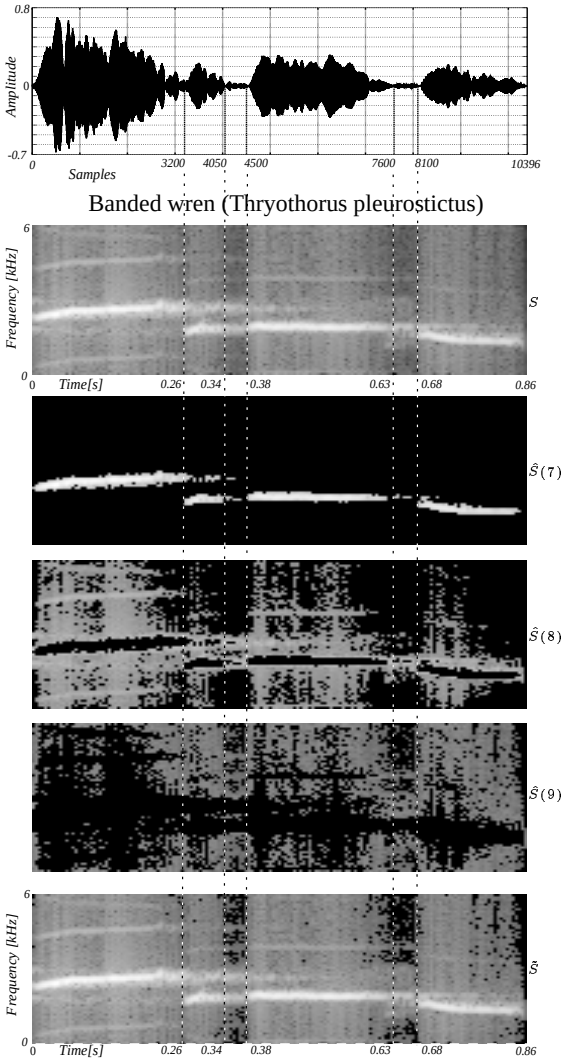

Figure 4: *Processing result (non-speech).*

The three quite clear tones (see formants in Figure 3) of the birds song are masked very well in $\hat{S}(7)$. Hearing impression yields no notable difference to the original signal. Obviously, the low-energy components of the signal are filtered. Note the potential for signal compression with $\hat{S}(7)$ defining the frequency-range. As in the results presented before, the next mask $\hat{S}(8)$ complements its predecessor, containing homogeneous regions of lower amplitude. Again, the envelope of the original sound is kept. $\hat{S}(9)$ contains only background-noise. Due to the wide coverage of signal characteristics in the masks used in the composition $\tilde{S}$, we achieve almost a reconstruction of the original signal.

5 CONCLUSIONS

A reduced signal-representation using spectro-temporal components is generated from the temporal segmentation abilities of a pulsed coupled neural network. This lead to an approach for the enhancement of speech and other audio signals, based on the composition of selected masked sonagram representations. Results imply that phonetically irrelevant information can also be separated. The reduced *masked sonagram representation*, that uses spectro-temporal components, is promoted by redundancy in the signal-representation.

A quantitative measure of segment quality (e.g. employing the mutual information of S and a particular $\hat{S}$) could also be integrated in the process to choose significant segments automatically, and to then optimally compose an enhanced signal from its masked components. An on-line adaption of the PCNN-parameters, with regard to the measured quantity, could be used to automate the enhancement process itself. Next to supporting the recognition of speech, the visual perceivability of the approach offers a different quality of visual speech understanding for hearing-impaired.

6 REFERENCES

[1] E. A. Wan and A. T. Nelson, "Networks for speech enhancement," in *Handbook of Neural Networks for Speech Processing*, Shigeru Katagiri, Ed., pp. 541–541. Artech House, 2000.

[2] J. M. Zaruda, Ed., *Special Issue on Pulse Coupled Neural Networks*, vol. 10 of *IEEE Transactions on Neural Networks*, IEEE Neural Networks Council, May 1999.

[3] W. Maass and C. M. Bishop, Eds., *Pulsed Neural Networks*, MIT Press, 1999.

[4] M. Garzon, *Models of Massive Parallelism: Analysis of Cellular Automata and Neural Networks*, Texts in theoretical computer science. Springer-Verlag, 1995.

[5] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *Journal of Physiology*, vol. 117, pp. 500–544, 1952.

[6] R. Eckhorn, H. J. Reitboeck, M. Arndt, and P. Dicke, "Feature linking via synchronization among distributed assemblies: Simulations of results from cat visual cortex," *Neural Computation*, vol. 2, pp. 293–307, 1990.

[7] M. P. Schamschula, J. L. Johnson, and R. Inguva, "Image processing with pulse coupled neural networks," in *Proc. of the 2nd International Forum on Multimedia and Image Processing, World Automation Congress*, Maui, Hawaii, 2000.

[8] J. L. Johnson and M. L. Padgett, "PCNN Models and Applications," *IEEE Transactions on Neural Networks*, 1999.

[9] J. L. Flanagan, *Speech Analysis, Synthesis, and Perception*, Springer-Verlag, New York, 1972.

[10] R. J. J. H. van Son, D. Binnenpoorte, H. van den Heuvel, and L. C. W. Pols, "The IFA Corpus: a Phonemically Segmented Dutch 'Open Source' Speech Database," in *Proc. of Eurospeech 2001*, Aalborg, Denmark, Sept. 3-7 2001.

[11] J. M. Burt, "Use of a radio microphone array to study banded wren song interactions at the neighborhood level," *Journal of the Acoustical Society of America*, vol. 108, no. 5, pp. 2583, 2000.