THE NEUROMORPHIC ENGINEER

VOLUME 3 NUMBER 2 MARCH 2007

In this issue:

Brain-inspired auditory processor and the *Artificial Brain* project

The Korean Brain Neuroinformatics Research Program has two goals: to understand information processing mechanisms in biological brains and to develop intelligent machines with human-like functions based on these mechanisms. We are now developing an integrated hardware and software platform for brain-like intelligent systems called the *Artificial Brain*. It has two microphones, two cameras, and one speaker, looks like a human head, and has the functions of vision, audition, inference, and behavior (see Figure 1).

The sensory modules receive audio and video signals from the environment, and perform source localization, signal enhancement, feature extraction, and user recognition in the forward 'path'. In the backward path, top-down attention is performed, greatly improving the recognition performance of real-world noisy speech and occluded patterns. The fusion of audio and visual signals for lip-reading is also influenced by this path.

The inference module has a recurrent architecture with internal states to implement human-like emotion and self-esteem. Also, we would like the *Artificial Brain* to eventually have the abilities to perform user modeling and active learning, as well as to be able to ask the right questions both to the right people and to other *Artificial Brains*.

The output module, in addition to the head motion, generates human-like behavior with synthesized speech and facial representation for 'machine emotion'. It also provides computer-based services for users.

The *Artificial Brain* may be trained to work on specific applications, and the *OfficeMate* is our choice of application test-bed. Similar to office secretaries, the *Of-*

***Lee,** continued p. 12*

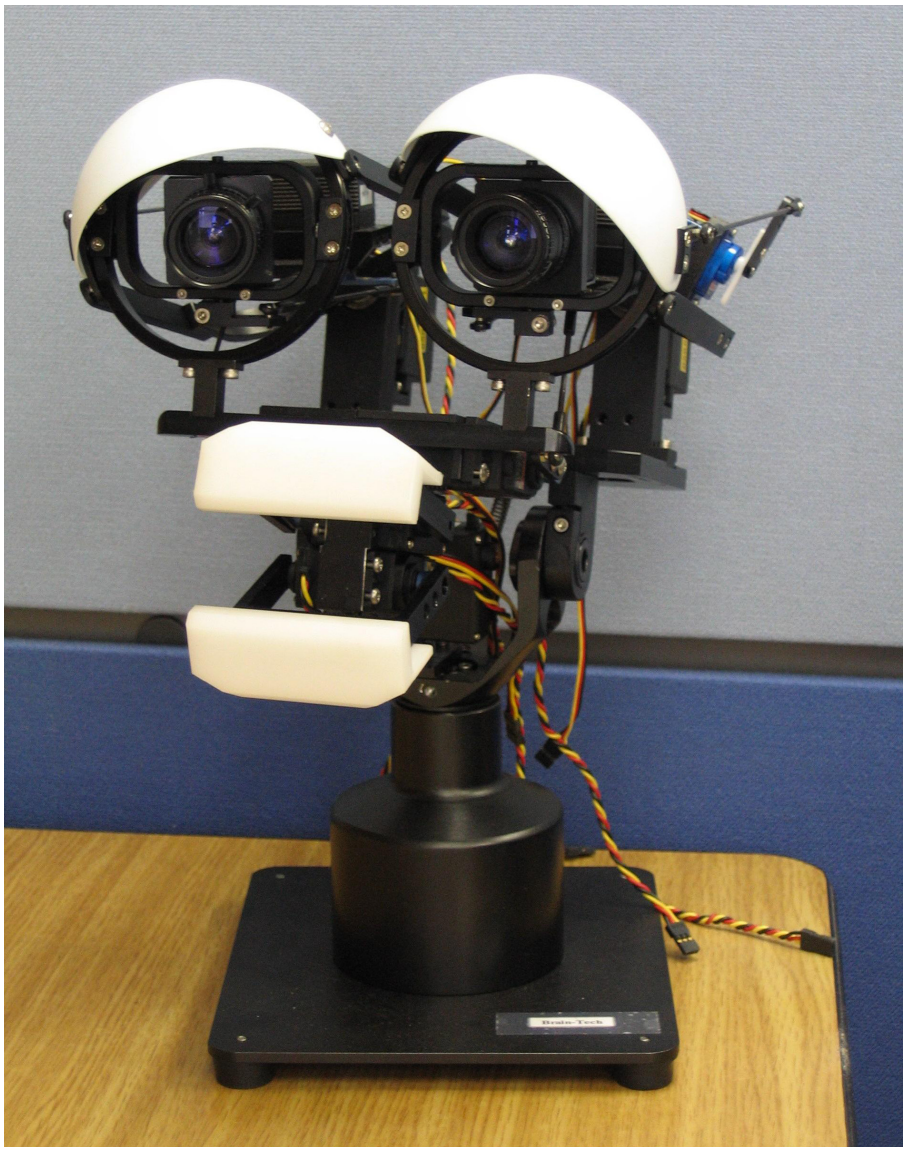


Figure 1. An Artificial Brain with two eyes (cameras), two ears (microphones), and one mouth (speaker). It can interact with humans via intelligent functions such as sound localization, speech enhancement, user and emotion recognition with speech and face images, user modeling, and active learning of new things.

THE NEUROMORPHIC ENGINEER

is published by the

INSTITUTE OF NEUROMORPHIC ENGINEERING

Editor
Sunny Bains
Imperial College London
sunny@sunnybains.com

Editorial Assistant
Dylan Banks

Editorial Board
David Balya
Avis Cohen
Tobi Delbrück
Ralph Etienne-Cummings
Timothy Horiuchi
Auke Ijspeert
Giacomo Indiveri
Shih-Chii Liu
Jonathan Tapson
André van Schaik
Leslie Smith

This material is based upon work supported by the National Science Foundation under Grant No. IBN-0129928. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The Institute of Neuromorphic Engineering
Institute for Systems Research
AV Williams Bldg.
University of Maryland
College Park, MD 20742

http://www.ine-web.org

EDITORIAL

What to feed the senses?

I've been thinking about getting information into the human brain. A feature for *Wired* magazine I finished recently discusses ways to feed in new (and potentially strange) kinds of information through the periphery: so, for instance, I got to try a 'tongue display' where images acquired by a camera on my forehead were fed into an array of pixels passing small currents through my tongue. I also got to use a haptic vest designed to help pilots understand their spatial orientation. I wanted to understand how much information the senses could handle (no point in supplying more), how much the brain could handle (without confusion), and how signals should be encoded.

I won't go into the details: you can check out the article[1] or look at my blog[2] if you're interested. But a number of interesting questions came up during my research: questions that the neuromorphic community may either be interested in or be able to help me with.

The first thing I found fascinating was how knowledge of sensory bandwidth—which seems like it should be crucial to all engineering in this area—seemed very sketchy. For instance, one recent paper[3] says the retina has about the same bandwidth as ethernet, 10Mb/s. But it doesn't seem to relate the image coming in through the eye to that being passed through to the brain. In particular, nowhere in the paper does it suggest the retina might be doing some kind of *compression,* which seemed to me like an important issue (even if only to address and dismiss). Also, I practically begged a rearcher in tactile displays at the University of Madison (who worked on both tongue and fingertip displays) to tell me where I could get figures for tactile bandwidth. In his opinion, it was a meaningless question. I certainly couldn't find any useful literature on the subject myself: not for this or any of the other senses I was looking at.

The other main issue that started to intrigue me was attention. I know this makes me slow on the uptake, since this has been a 'hot topic' for a long time, but I had no particular reason to be deeply interested until now.

Specifically, I've become intrigued by two things: how attention is split up within a particular sense, and among the senses. My interest came from the fact that the tongue-display system I used, which was intended to help people with macular degeneration, felt very 'visual'. My memories of using the device (blindfolded) are not of feeling sensations in the tongue, but instead of seeing a black-and-white low-resolution world. I'm told that this 'visual' feeling is probably due to the fact that the information from the information is feeding into the visual part of the extrastriate cortex, the bit involved with mental imagery. Which begs a question: just how much imagery can a person handle, and does it matter where that imagery comes from?

An interesting development that relates a little to this is the work of some military researchers investigating the advisability of feeding different images—say a close-up of a sniper and a view of the whole building or scene—into left and right eyes. You can read the work yourself, but the bottom line is that it doesn't work well: performance and reaction times drop because the brain doesn't seem to be able to take it all in.[4]

There are many theories of attention, of course, but one presented by a colleague of mine here at Imperial College London recently seemed very persuasive.[5] (Even though he used the 'C' word, consciousness, when he described it.) He presented new work neuromodelling something called the global workspace theory to show how different sensory inputs can compete with each other to produce psychological phenomena that we know take place, and which seems to have biological plausibility. Of course, it's still far from answering the engineering question, 'Exactly how much can we usefully put in?'

One last thing I've been wondering about (to no avail so far), is whether the form of an incoming signal matters as much as its source. Specifically, if 'visual' information (images of remote objects, rather than those on the 2D periphery of the body) comes in through the tongue, how is the bit of the brain that deals with the tongue equipped to decipher it? Does it get help from some bit of the visual cortex (perhaps a higher-level part) or does it just figure out how to process images?

Any answers or leads would be much appreciated: and I promise to share them!

Sunny Bains
Editor, The Neuromorphic Engineer
http://www.sunnybains.com

References
1. Sunny Bains, *Mixed feelings,* **Wired 15.4,** April 2007.
2. http://www.sunnybains.com/blog
3. Kristin Koch et. al., *How Much the Eye Tells the Brain,* **Current Biology 16 (14), 2006.**
4. David Curry et al., *Dichoptic image fusion in human vision system,* **Proc. SPIE 6224,** 2006.
4. Murray Shanahan, *A Spiking Neuron Model of Cortical Broadcast and Competition,* **Consciousness and Cognition,** 2007 (in press).

Supervised learning in spiking neural networks

Spiking neural networks (SNN)[1-5] exhibit interesting properties that make them particularly suitable for applications that require fast and efficient computation and where the timing of input/output signals carries important information. However, the use of such networks in practical, goal-oriented applications has long been limited by the lack of appropriate supervised-learning methods. Recently, several approaches for learning in SNNs have been proposed.[3] Here, we will focus on one called ReSuMe[2,6] (remote supervised method) that corresponds to the Widrow-Hoff rule and is well known from traditional artificial neural networks. ReSuMe takes advantage of spike-based plasticity mechanisms similar to spike-timing dependent plasticity (STDP).[1,6] Its learning rule is defined by the equation below:

$$\frac{\mathrm{d}}{\mathrm{d}t}w_{ki}(t) = \left[S^d(t) - S^o(t)\right]\left[a + \int_0^\infty W(s)\, S^{in}(t-s)\,\mathrm{d}s\right],$$

where $S^d(t)$, $S^{in}(t)$ and $S^o(t)$ are the desired pre- and postsynaptic spike trains,[1] respectively. The constant a represents the so-called non-Hebbian contribution to the weight changes. The role of this parameter is to adjust the average strength of the synaptic inputs so as to impose on a neuron the desired level of activity (desired mean firing rate). The function $W(s)$ is known as a learning window[1] and, in ReSuMe, its shapes of are similar to those used in STDP models. The parameter s is a time delay between the correlated spikes. (For a detailed introduction to ReSuMe, please see Reference 6.)

It has been demonstrated that ReSuMe enables effective learning of complex temporal and spatio-temporal spike patterns with a given accuracy (see Figure 1) and that the method enables us to impose desired input/output properties on the networks.[2,10] Contrary to most existing supervised learning methods in SNN, ReSuMe is independent of the spiking neuron models and can be effectively applied to the broad class of spiking neurons.[2,7] Convergence of the ReSuMe learning process has been formally proved for some classes of the learning scenarios.[7]

In Reference 10 we demonstrated the generalization properties of the spiking neurons trained with ReSuMe. It was also shown that SNNs are able to perform function approximation tasks. Moreover, we demonstrated that, by appropriately setting the learning rule parameters, networks can be trained to reproduce desired spiking patterns $S^d(t)$ with a controllable time lag Δt, such that the reproduced signal $S^o(t) \cong S^d(t - \Delta t)$ (unpublished results). This property has very important outcomes for the possible applications of ReSuMe: e.g. in prediction tasks, where SNN-based adaptive models could predict the behaviour of reference objects in on-line mode.

Especially promising applications of SNN are in neuroprostheses for human patients with the dysfunctions of the visual, auditory, or neuro-muscular systems. Our initial simulations in this area point out the suitability of ReSuMe as a training method for SNN-based neurocontrollers in movement generation and control tasks.[8,9,11]

Real-life applications of SNN require efficient hardware implementations of the spiking models and the learning methods. Recently, ReSuMe was tested on an FPGA platform:[4] the implemented system demonstrated fast learning convergence and the stability of the optimal solutions obtained. Due to its very fast processing ability, the system is able to meet the time restrictions of many real-time tasks.

Filip Ponulak
Inst. of Control and Information Eng.
Posnan University of Technology
Posnan, Poland
E-mail: Filip.Ponulak@put.poznan.pl
http://d1.cie.put.poznan.pl/~fp

References

1. W. Gerstner and W. Kistler, **Spiking Neuron Models. Single Neurons, Populations, Plasticity,** Cambridge University Press, 2002.
2. A. Kasinski and F. Ponulak, *Experimental Demonstration of Learning Properties of a New Supervised Learning Method for the Spiking Neural Networks,* **15th Int'l Conf. on Artificial Neural Networks: Biological Inspirations 3696,** pp. 145–153, Warsaw, 2005.
3. A. Kasinski and F. Ponulak, *Comparison of Supervised Learning Methods for Spike Time Coding in Spiking Neural Networks,* **Int'l J. of App. Math. and Comp. Sci. 16** (1), pp. 101–113, 2006.
4. M. Kraft, A. Kasinski, and F. Ponulak, *Design of the spiking neuron having learning capabilities based on FPGA circuits,* **Third Int'l IFAC Workshop on Discrete-Event System Design,** pp. 301-306, Rydzyna, 2006.
5. W. Maass, *Networks of spiking neurons: The third generation of neural network models,* **Neural Networks 10** (9), pp. 1659-1671, 1997.
6. F. Ponulak, **ReSuMe—new supervised learning method for Spiking Neural Networks,** Technical Report, 2005.http://d1.cie.put.poznan.pl/~fp/.
7. F. Ponulak, **ReSuMe—Proof of convergence,** Technical Report, 2006. http://d1.cie.put.poznan.pl/~fp/.
8. F. Ponulak, D. Belter, and A. Kasinski, *Adaptive Central Pattern Generator based on Spiking Neural Networks, Dynamical principles for neuroscience and intelligent biomimetic devices,* **EPFL LATSIS Symp.,** pp. 121–122, Lausanne, 2006.
9. F. Ponulak and A. Kasinski, *A novel approach towards movement control with Spiking Neural Networks,* **Third Int'l Symp. on Adaptive Motion in Animals and Machines,** Ilmenau, 2005. (Abstract)
10. F. Ponulak and A. Kasinski, *Generalization Properties of SNN Trained with ReSuMe,* **Euro. Symp. on Artificial Neural Networks**, pp. 623–629, Bruges, 2006.
11. F. Ponulak and A. Kasinski, *ReSuMe learning method for Spiking Neural Networks dedicated to neuroprostheses control: Dynamical principles for neuroscience and intelligent biomimetic devices,* **EPFL LATSIS Symp.,** pp. 119–120, Lausanne, 2006.

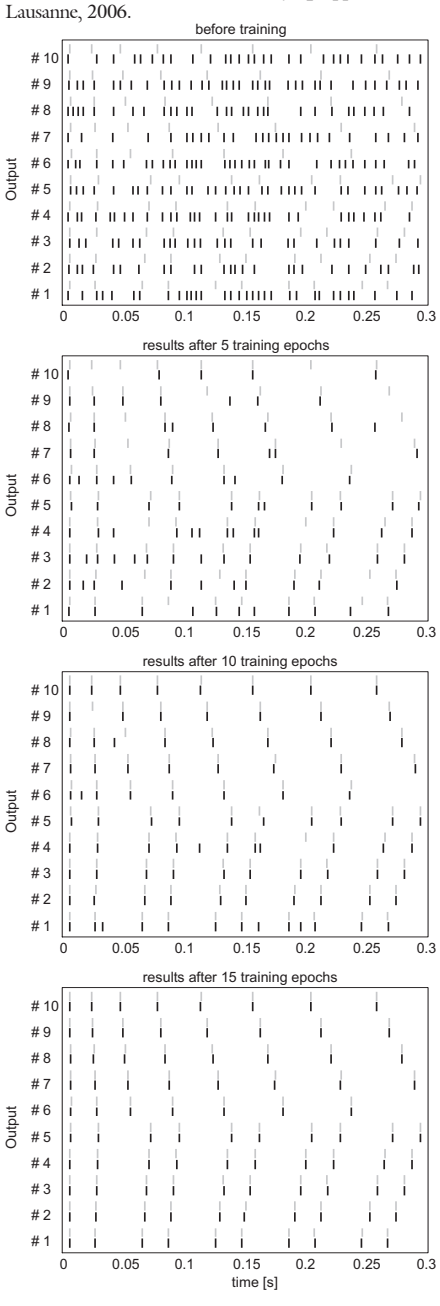


Figure 1. ReSuMe is used to train a spiking neural network to store and recall an exemplary target spatio-temporal pattern of spikes (gray bars). The spike-trains produced at the network outputs (black bars) before the training (A) and after 5, 10 or 15 learning epochs (B,C,D, respectively) are shown. After 15 learning epochs the target pattern is almost perfectly reproduced, with the correlation equal to 0.998.

Dealing with unexpected words

Insperata accidunt magis saepe quam quae speres, i.e. things you do not expect happen more often than things you do expect, warns Plautus (circa 200 BC). Most readers would agree with Plautus that surprising sensory input data could be important since they could represent a new danger or new opportunity. A hypothesized cognitive process involved in the processing of such inputs is illustrated in Figure 1.

In machine recognition, low-probability items are unlikely to be recognized. For example, in the automatic recognition of speech (ASR), the linguistic message in speech data X is coded in a sequence of speech sounds (phonemes) Q. Substrings of phonemes represent words, sequences of words form phrases. A typical ASR attempts to find the linguistic message in the phrase. This process relies heavily on prior knowledge in text-derived language model and pronunciation lexicon. Unexpected lexical items (words) in the phrase are typically replaced by acoustically acceptable in-vocabulary items.[1]

Our laboratory is working on identification and description of low-probability words as a part of the large multinational DIRAC project (Detection and Identification of Rare Audio-Visual Cues), recently awarded by the European Commission. Principles of our approach are briefly described below.

To emulate the cognitive process shown in Figure 1, the contemporary ASR could provide the predictive information stream. Next we need to estimate similar information without the heavy use of prior knowledge. For the estimation of context-constrained and context-unconstrained phoneme posterior probabilities, we have used a continuous digit recognizer based on a hybrid Hidden-Markov-Model Neural-Network (HMM-NN) technique,[1] shown schematically in Figure 2. First, the context-unconstrained phoneme probabilities are estimated. These are subsequently used in the search for the most likely stochastic model of the input utterance. A by-product of this search is a number of context-constrained phoneme probabilities.[2]

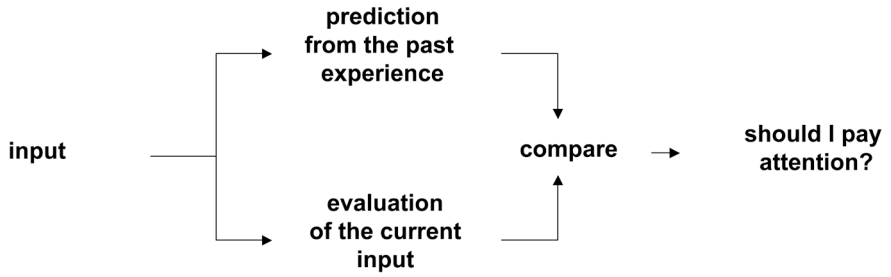


Figure 1. Hypothesized process for the discovery of unexpected items. The sensory input triggers a predictive process in the upper path that uses top-down knowledge from the past experience and generates predicted components of the scene. In parallel, the scene components are also estimated directly (i.e. without the use of the top-down knowledge) from the input. A comparison between the two sets of components may indicate an unexpected item.

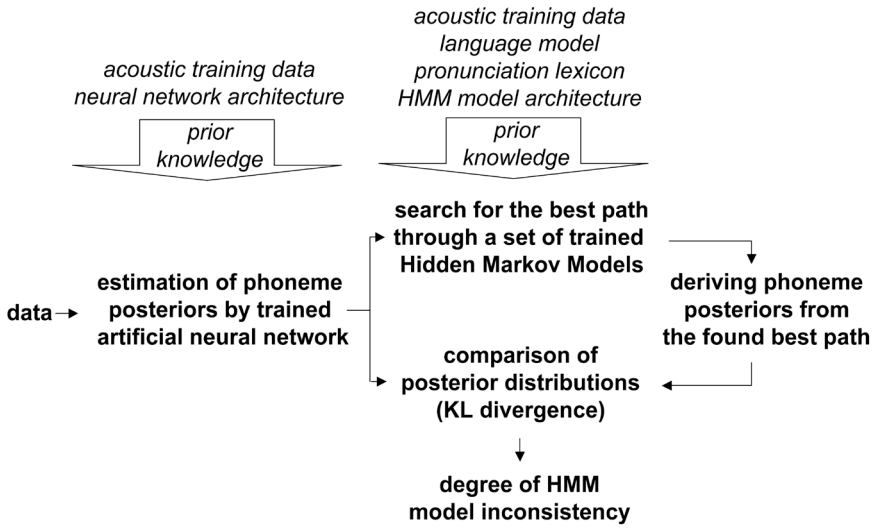


Figure 2. Discovery of out-of-vocabulary words using the hybrid HMM-NN ASR system, in which the out-of-context posterior probabilities estimated by the artificial neural network (ANN) are also directly used in the constrained search for the best model sequence.[2]

The basic principles of deriving the context-unconstrained posterior probabilities of phonemes are illustrated in Figures 3 and 4. A feed-forward artificial neural network is trained on phoneme-labelled speech data and estimates unconstrained posterior probability density function $p_i(Q|X)$.[3] This uses as an input a segment x_i of the data X that carries the local information about the identity of the underlying phoneme at the instant i. This segment is projected on 448 time-spectral basis. As seen in the middle part of Figure 5, the estimate from the NN can be different from the estimate from the context-constrained stream since it is not dependent on the constraints L.

The context-unconstrained phoneme probabilities can be used in a search for the most likely Hidden Markov Model (HMM) sequence that could have produced the given speech phrase. As a side product, the HMM can also yield, for any given instant i of the message, its estimates of posterior probabilities of the hypothesized phonemes $p_i(Q|X,L)$ 'corrected' by a set of constraints L implied by the training-speech data, model architecture, pronunciation lexicon, and the applied language model.[4] When it encounters an unknown item in the phoneme string (e.g. the word 'three' in Figure 5), it assumes it is one of the well known items. Note that these 'in context' posterior probabilities, even when wrong, are estimated with high confidence.

An example of a typical result[4] is shown in Figure 5. As seen in the lower part of the figure, an inconsistency between these two information streams could indicate an unexpected out-of-vocabulary word.

Being able to identify which words are not in the lexicon of the recognizer, and being able to provide an estimate of their pronunciation, may allow for inclusion of

***Hermansky,** continued p. 5*

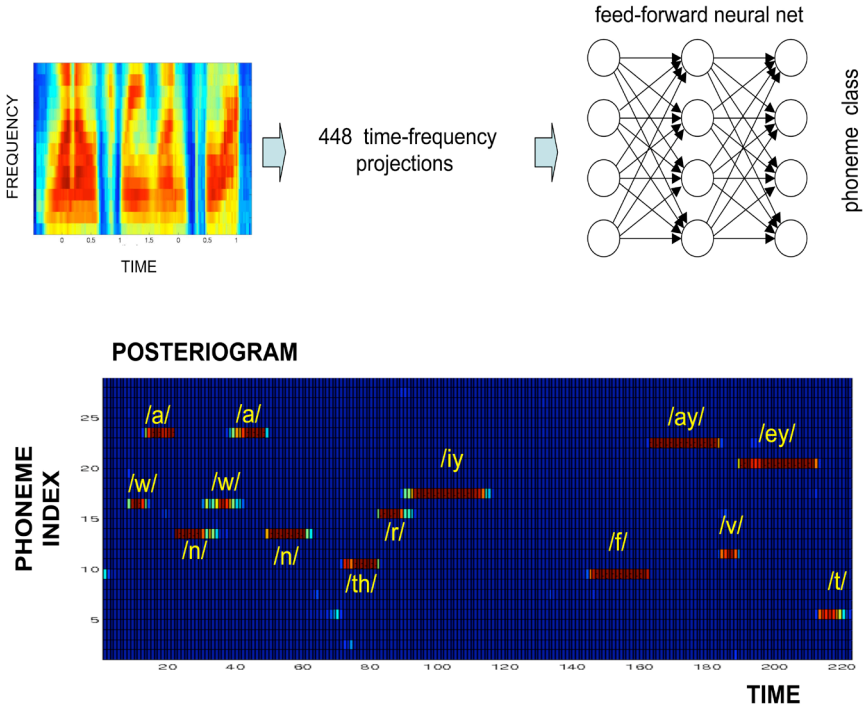


Figure 3. Posterior probabilities of phonemes estimated by an HMM-based system (the upper part of the Figure), and by ANN (the middle part of the Figure). In this example, the HMM model inconsistency was introduced by removing the word three from the recognizer vocabulary. The correct phoneme sequence for the word three is misrepresented in the HMM-derived posteriogram (replaced by a sequence /z/iy//r//oh/ of the in-vocabulary word zero). The ANN-derived probabilities indicate in this case the correct sequence /th//r//iy/ for the out-of-vocabulary word three. Comparison of the respective posterior probability density functions by evaluating their relative entropy (also known as KL divergence): its running average, evaluated over 100 ms time intervals, is shown in the lower part of the figure. This indicates HMM model inconsistency in the neighbourhood of the out-of-vocabulary word there (see Reference 4 for more details).

***Unexpected words,** continued from p. 4*

these new words in the pronunciation dictionary, thus leading to an ASR system that would be able to improve its performance as being it is used over time, i.e. that is able to learn. However, the inconsistency between in-context and out-of-context probability streams need not indicate the presence of unexpected lexical item but could indicate other inadequacies of the model. Further, this inconsistency might also indicate corrupted input data if the in-context probability estimation using the prior L yields more reliable estimate than the unconstrained out-of-context stream. Thus, providing a measure of confidence in the estimates from both streams would be desirable when corrupted input is a possibility.

Hynek Hermansky
IDIAP Research Institute
Swiss Federal Institute of Technology
Lausanne, Switzerland
Email: hynek.hermansky@idiap.ch

References
1. H. Hermansky and N. Morgan, *Automatic Speech Recognition*, in **Encyclopedia of Cognitive Science,** L. Nadel, Ed., Nature Publishing Group, Macmilian Publishers, 2002.
2. Bourlard, H. and Morgan, N., **Connectionist Speech Recognition—A Hybrid Approach**, Kluwer Academic

***Hermansky,** continued p. 10*

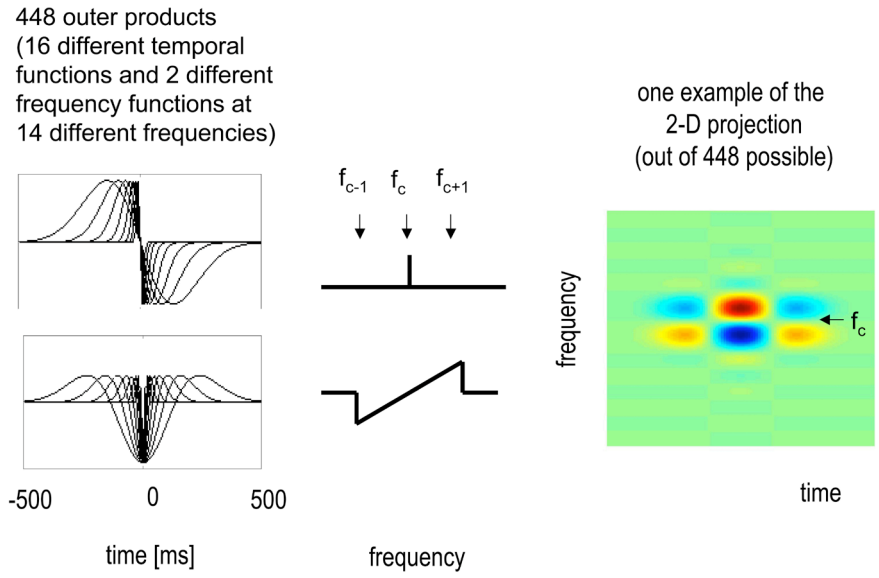


Figure 4. Illustration of the technique for obtaining a reliable estimate of posterior probability density functions $p_i(Q|X)$ without the use of top-down constraints L. The short-term critical-band spectrogram (left part of the figure) is derived by weighted summation of appropriate components of the short-term spectrum of speech. A segment of this spectrogram is projected on 448 different time-frequency bases (shown in Figure 3), centred at the time instant i, yielding a 448 point vector that forms the input to the MLP neural net, trained on about 2 hours of hand-labelled telephone-quality speech to estimate a vector of posterior probabilities $p_i(Q|X)$. A set of $p_i(Q|X)$ for all time instants forms the so-called posteriogram, shown for the utterance one-one-three-five-eight in the lower part of the figure. Higher posterior probabilities are indicated by warmer colors (see Reference 5 for more details).

Embedded vision system for real-time applications

There is an increasing demand for low-power, low-cost, real-time vision systems able to perform a reliable analysis of a visual scene: especially in environments where the lighting is not controlled. In the automotive industry, for instance, there are many potential applications including lane-departure warnings, seat-occupancy detection, blind-angle monitoring and pedestrian detection. But there are multiple constraints involved in embedding a vision system in a vehicle. First, the automotive industry has stringent requirements in terms of cost. Second, a vision system in a moving vehicle will experience sudden changes in illumination level and wide intra-scene dynamic range: this imposes severe constraints on the sensor characteristics and the optical design. Finally, the diversity of environments and situations, and the need for a fast reaction time make algorithm development a challenging part of the work.

The approach we have taken to solving these multiple requirements consists in moving part of the image processing to the sensor itself. This allows the extraction of robust image features independent of the illumination level and variation, and limits data transmission to the features required to perform a given task. The vision sensors developed at CSEM[1,2] perform the computation of the contrast magnitude and direction of local image features at the pixel level by taking spatial derivatives at each pixel. These derivatives are multiplied by a global steering function varying in time, resulting in a sinusoidal signal whose amplitude and phase represent, respectively, the contrast magnitude and direction.

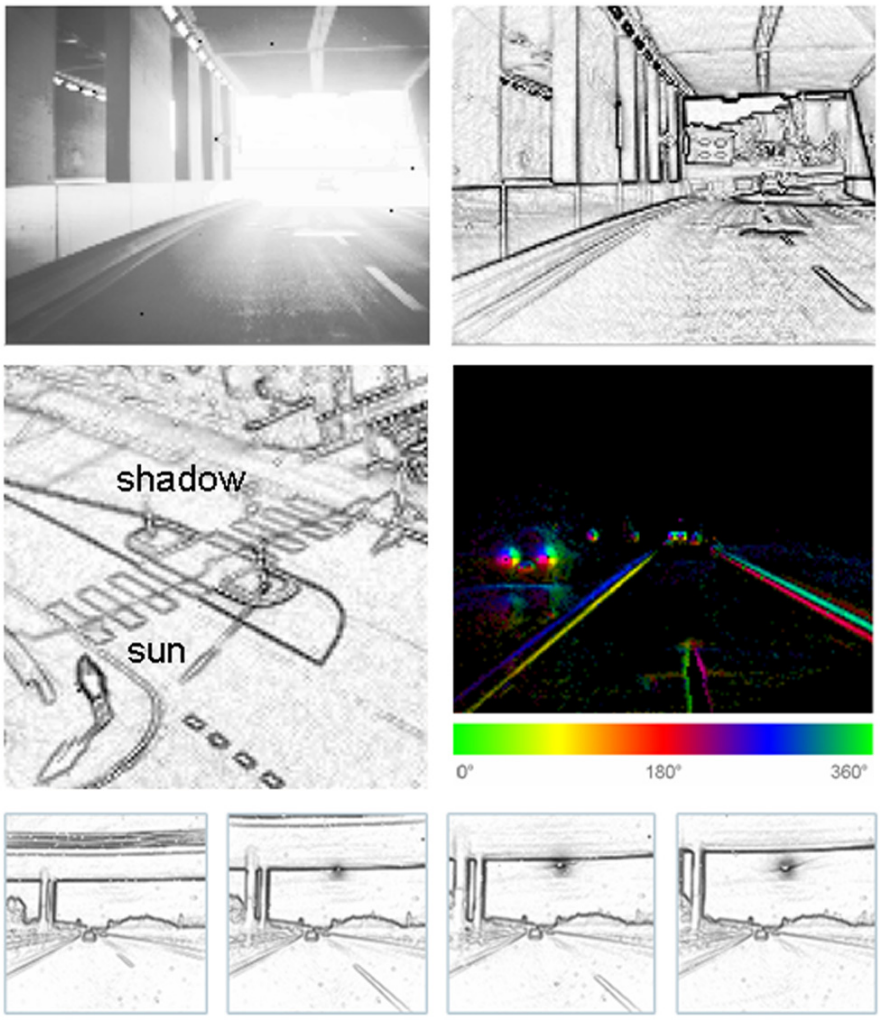


Figure 1. Shown is the gray-level image (top left) and contrast magnitude representation (top right) close to a tunnel exit. Middle left is the contrast magnitude representation with a transition between a sunny area and a shadowed area across the pedestrian crossway. Middle right shows the contrast direction representation on a road at night, with cars coming in the opposite direction. Here the representation is color-encoded. The bottom row shows the contrast magnitude with the sun entering in the field of view.

The contrast representation derived in the vision sensor is equivalent to normalizing the spatial gradient magnitude with the local intensity. Unlike the spatial gradient, the contrast representation does not depend on illumination strength, thus introducing considerable advantages for the interpretation of scenes. Furthermore, information is dispatched by decreasing order of contrast magnitude, thus prioritizing pixels where contrast magnitude is strong: these are usually sparse in natural images.[3] This mechanism allows reducing the amount of data dispatched by the sensor. Figure 1 illustrates the high intra-scene dynamic range of the vision sensor and its ability to discard illumination.

A compact and low-power platform, called *Devise*, has been developed to demonstrate the efficiency of this approach to implement low-power real-time vision systems. The platform, shown on Figure 2, embeds a vision sensor,[2] a BlackFin BF 533 processor, memory, and communication interfaces. An Ethernet interface enables easy connection to a PC, allowing visualization of raw data in real time and easing the development and debugging of new algorithms. Once an application has been developed and migrated to the BlackFin processor, a low-data-rate radio-frequency link is available that can be used, for instance, to communicate between different nodes in a network of such platforms.

In the last few years, we have made a continuing effort to develop software that exploits the contrast information delivered by our vision sensors to analyze visual scenes in natural environments. Development has been focused in two areas: automotive, as mentioned previously, and surveillance. The main function of our 'driver assistant' algorithm, for instance, is to detect the road markings so that the position of the vehicle on the road is known at all times and the driver can be warned if they leave their lane unintentionally. Each road marking—con-

Ruëdi and Grenet, *continued p. 7*

Figure 2. Shown left are the platform components: sensor board, processing board, battery, optic, and case. Right, the vision system can be seen mounted in a car behind the rear-view mirror for live lane-departure warnings.

Embedded vision, *continued from p. 6*

sisting of two edges with high contrast magnitude and opposite contrast directions—is detected and tracked in a restricted area that is continuously adapted to the last detected position. Continuous and dashed markings are differentiated. The vanishing point is extracted, the variations of which give useful gyroscopic information (tilt and yaw angles). A Kalman filter supervises the system and gives robustness to the detection (e.g. when markings are temporarily missing). The system also estimates the illumination level and road curvature by fitting the markings points with a clothoid equation, allowing it to appropriately control the headlights.

This algorithm, implemented in the BlackFin processor, works robustly at 25 frames per second in varying conditions such as night, sun in the field of view, and roads with poor quality markings. For demonstration purposes, detection results (mark position and type, road curvature, light level, etc.) are sent via the low-data-rate radio-frequency link to a cellular phone that displays a synthetic view of the road in real time (see Figure 3).

This work demonstrates that moving some of the image processing to the sensor itself is a solution to implement real-time low-power and low-cost vision systems able to function robustly in uncontrolled environments.

Pierre-François Rüedi and Eric Grenet
CSEM S.A.
Neuchâtel, Switzerland
E-mail: pfr@csem.ch, egt@csem.ch

References

1. M. Barbaro, P.-Y. Burgi, A. Mortara, P. Nussbaum and F. Heitger, *A 100×100 pixel silicon retina for gradient extraction with steering filter capabilities and temporal output coding*, **IEEE J. Solid-State Circuits 37** (2), pp. 160-172, 2002.
2. P.-F. Rüedi, P. Heim, F. Kaess, E. Grenet, F. Heitger, P.-Y. Burgi, S. Gyger, P. Nussbaum, *A 128×128 Pixels 120 dB Dynamic Range Vision Sensor Chip for Image Contrast and Orientation Extraction*, **IEEE J. Solid-State Circuits 38** (12), pp. 2306-2317, Dec. 2003.
3. D. J. Field, *What is the goal of sensory coding?*, **Neural Computation 6**, pp. 559–601, 1994.
4. E. Grenet, *Embedded High Dynamic Vision System For Real-Time Driving Assistance*, **TRANSFAC '06**, San Sebastian, Spain, p. 120, October 2006.

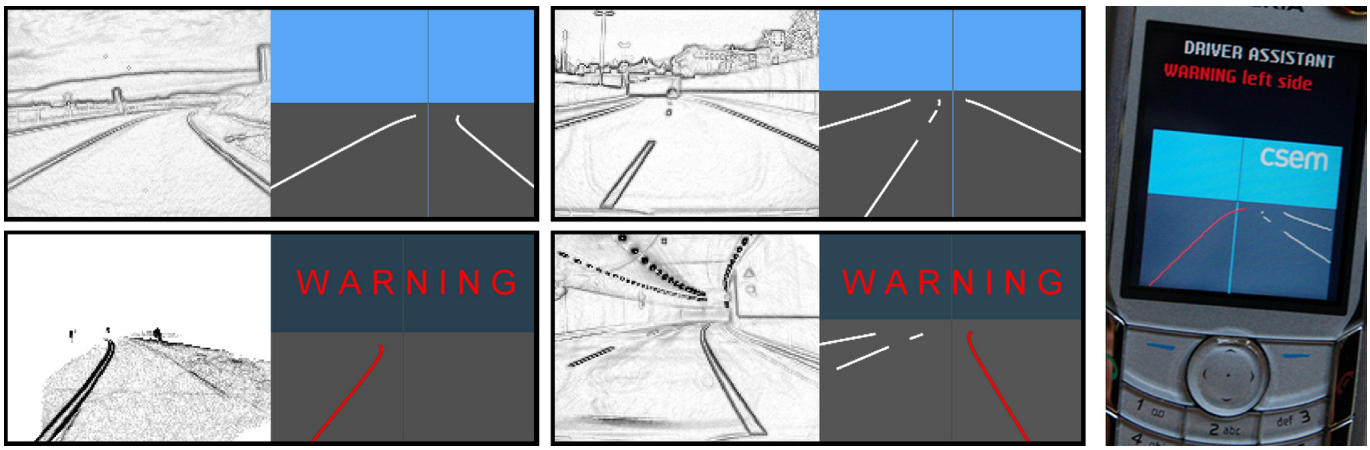


Figure3. Various road situations and their related symbolic representation. Shown are single-lane (top left) and a multi-lane curves (top middle) by day, a lane departure in a tunnel (bottom middle) and on a countryside road with single marking by night (bottom left). To the right is a real-time display and a warning on a cell phone.

Can spike-based speech recognition systems outperform conventional approaches?

The field of automatic speech recognition (ASR) has advanced far enough in the past decade to produce numerous commercial applications such as the speech-driven telephone customer service menus now deployed by many companies. Unfortunately, these and other state-of-the-art ASR systems still pale in comparison to human performance, particularly in the presence of noise. Researchers have long been aware of this discrepancy in performance and have often turned to biology seeking clues to the robustness of the human auditory system. As a matter of fact, the most commonly employed features for ASR applications are still the Mel frequency cepstral coefficients (MFCC), which mimic the logarithmic distribution of channels throughout the frequency of hearing as observed in the cochlea.

Nonetheless, today's ASR systems are designed with a window-based mindset using Hidden Markov Models (HMMs) and have little resemblance to neurobiological computation. As is well known, neurons in the brain use all-or-nothing action potentials to communicate timing information. These spike trains code sensory inputs and all levels of processing throughout the brain. Rather than being artifacts of biology, we believe that spike trains provide a key to the wonderful noise robustness of the auditory system and can be exploited in man-made machine recognition systems.

Recently, we proposed a spike-based classification scheme for simple acoustic signals that exploits the phase synchrony between the parallel streams of spike trains produced by the cochlea followed by a time-to-first-spike rank-order decoder for classification.[1] A more recent version of our system replaces the rank-order decoder with a spiking neural network for improved classification. Comparisons with a typical ASR engine show improved performance under the presence of noise. According to the results, spike firing times reveal a phase synchrony among tonotopically distributed auditory nerve fibers, which varies with the spectral properties of the input signal. Other researchers have proposed spike-based ASR systems but none have taken advantage of phase synchrony coding. We found out that the degree of such synchrony (DoS) constitutes a highly noise robust feature set for classification purposes by having little variation in response to changing noise levels.

***Uysal,** continued p. 9*

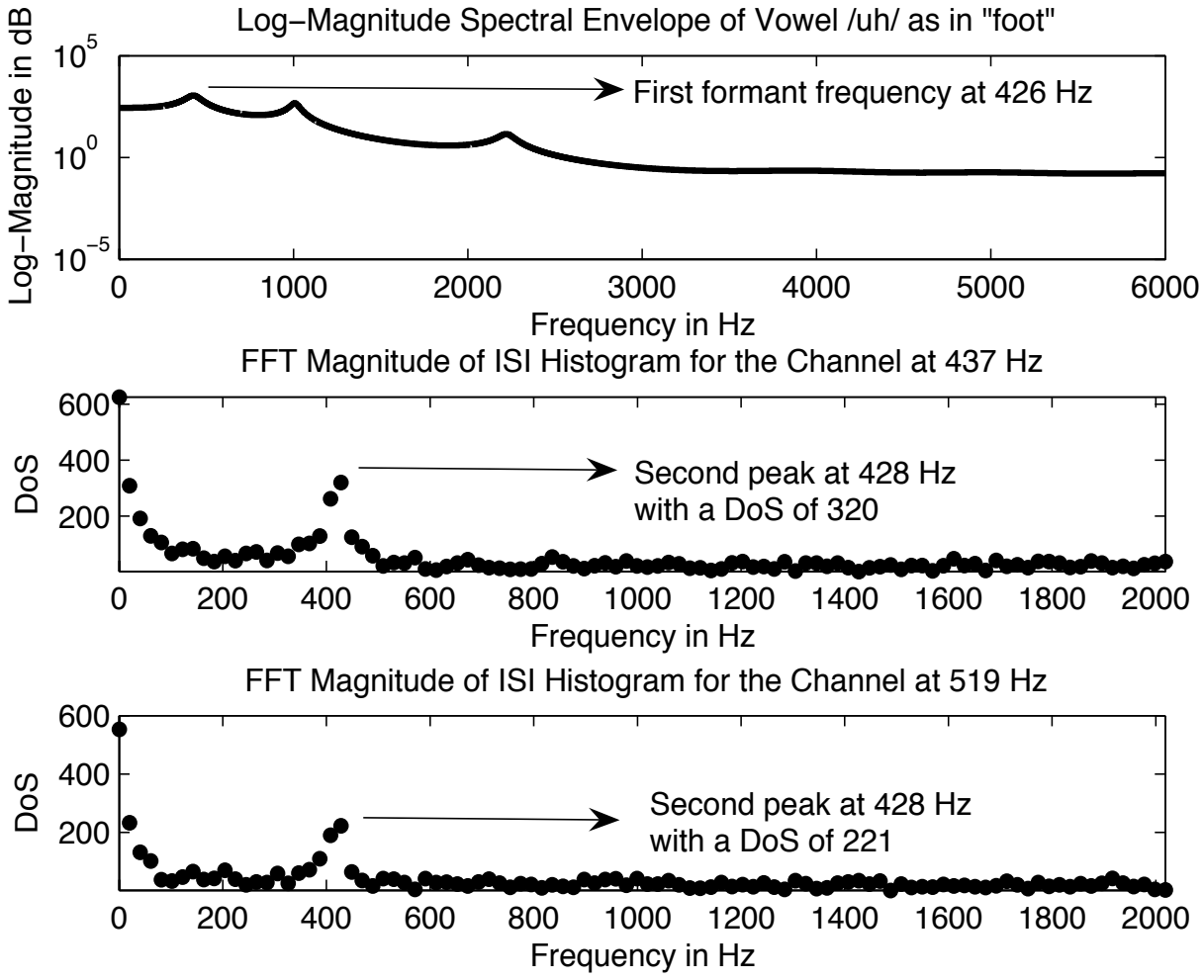


Figure 1. Log-magnitude spectral envelope for /uh/ and the corresponding degree of phase synchrony for two sets of hair cells centered at 437Hz and 519Hz (computed for a noisy utterance with 5dB SNR).

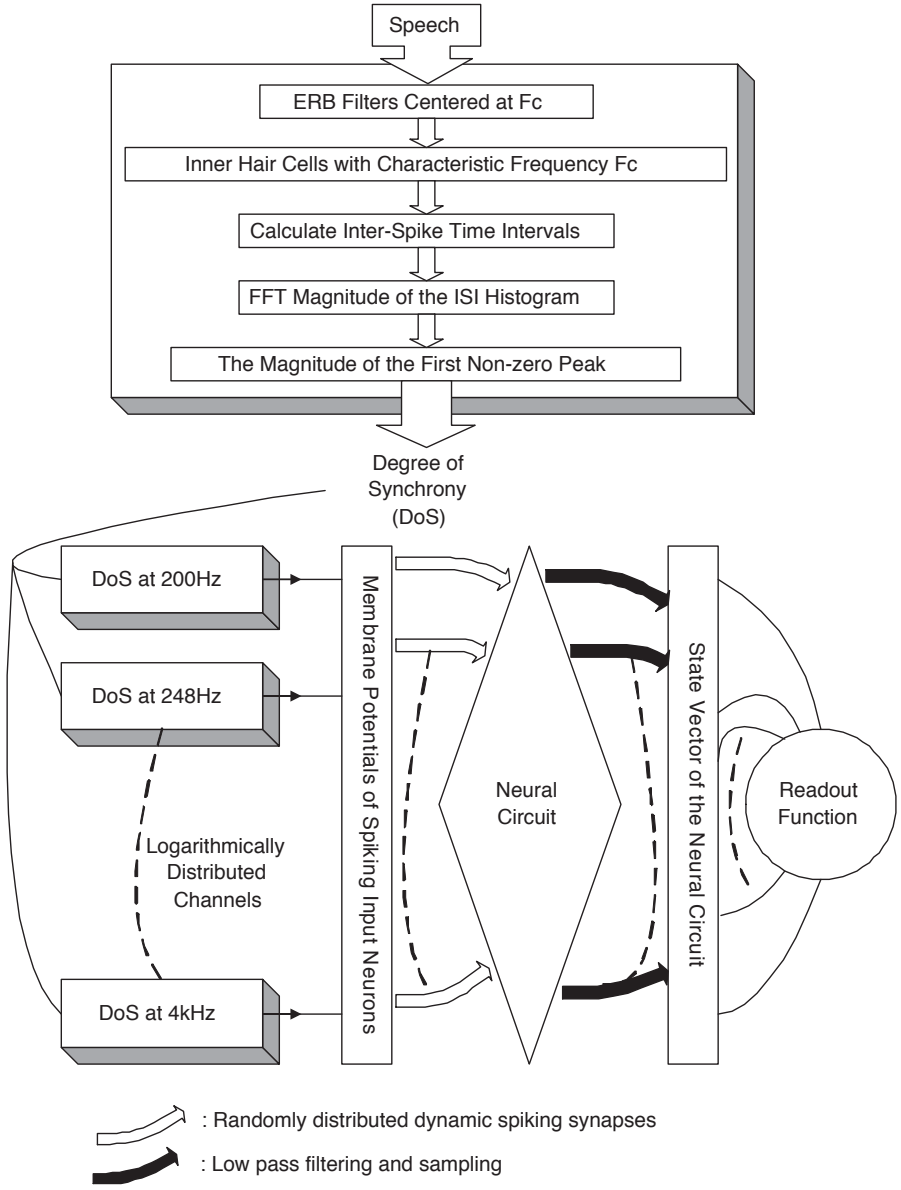


Figure 2. The overall spike-based classification. The degree of synchrony is extracted from spike trains generated in each individual cochlear channel. This feature set is then used with an LSM with supervised learning for classification.

Table 1.

SNR (dB) / System	25	10	5
MFCC - HMM	93%	86%	77%
Spike-based	92%	91%	89%

Spike-based speech, *continued from p. 8*

Spike-based classification architecture

The proposed system is composed of three main blocks: speech-to-spike conversion, feature extraction via phase synchrony coding, and classification via liquid state machine (LSM). For speech-to-spike conversion, we use an up-to-date cochlear simulation employing an improved inner hair cell model with auditory nonlinearities such as adaptation and temporal dynamics.[2] Human empirical data is used for various cochlear parameters, such as the distribution of the channels throughout the frequency of hearing.

For phase-synchrony coding, one has to look at the inter-spike time interval (ISI) histogram for each channel, which is defined as the total number of spikes falling within specified bins of time intervals. Our definition of the DoS for a particular channel is the magnitude of the first non-zero peak in the spectrum of its ISI histogram. As shown in Figure 1, even with a very noisy vowel input signal, the fibers with characteristic frequencies (437Hz) close to the first peak (426Hz) in the vowel's log-magnitude spectral plot are still able to phase lock very close to that particular frequency. They also have a higher DoS than other channels, such as the one shown in the bottom plot with a characteristic frequency (519Hz) further from the first formant peak.

Finally, for classification, the system employs an LSM with a randomly connected recurrent neural circuit.[3] The idea is to map the input vector to a higher dimension where the distance metric between prospective classes is larger. For our system, the input vector—which is comprised of the degrees of synchrony for each channel—is passed on to the neural circuit as the membrane potentials of input neurons that make dynamic spiking synapses with the circuit using spike-timing dependent plasticity. The state of the circuit is low-pass filtered and sampled to be associated with a target class (different types of vowels) by the help of a trainable readout function. Figure 2 shows the overall system design, as well as some of the important system parameters.

Results and discussion

We tested the algorithm on a noisy, multi-speaker, multi-gender vowel dataset. We compared the algorithm to a typical speech recognition engine employing the well-known MFCCs and an HMM. The percentage correct

Uysal, *continued p. 10*

Unexpected words, *continued from p. 5*

Publishers, 1994.
3. H. Bourlard, C. J. Wellekens, *Links between Markov Models and Multilayer Perceptrons,* **IEEE Conf. Neural Information Processing Systems**, 1988, Denver, CO, Ed. D. Touretzky, Morgan-Kaufmann Publishers, pp. 502-510, 1989.
4. H. Ketabdar and H. Hermansky, *Identifying and dealing with unexpected words using in-context and out-of-context posterior phoneme probabilities,* **IDIAP Research Report**, 2006.
5. H. Hermansky and P. Fousek, *Multi-resolution RASTA filtering for TANDEM-based ASR*, in **Proc. Interspeech 2005**, 2005.

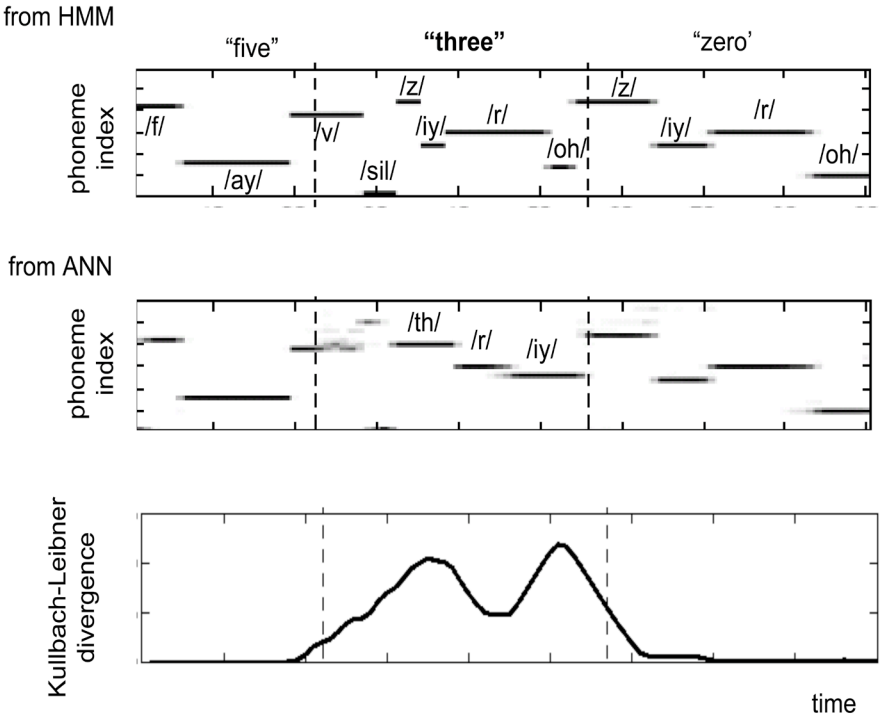


Figure 5. Shown are the time-frequency bases that attempt to emulate some very basic properties of auditory cortical receptive fields (e.g. Shamma). They are formed as outer products of first and second derivatives of truncated Gaussian functions of eight different widths in the time domain, and by summation and differentiation over three frequency components (three critical bands), centred at 14 different frequencies in the frequency domain (see Reference 4 for more details).

Spike-based speech, *continued from p. 9*

results are shown in Table 1.

At high signal-to-noise ratio (SNR) values, both systems perform comparably well, but the proposed system using phase synchrony coding is able to outperform the MFCC-HMM algorithm by 12% at 5dB SNR. In regards to the question raised in the title, though applied to a simplified domain, spike-based recognition is clearly more noise robust when compared to a conventional ASR system. This performance is mainly due to the phase synchrony maintaining capabilities of tonotopic neuron populations even under the presence of high amounts of noise.

Future work involves extrapolation of these findings to more complex signals and multi-syllable words by the help of relational networks as observed in the cortex.

Ismail Uysal, Harsha Sathyendra, and John G. Harris
Computational NeuroEngineering Lab
University of Florida
Gainesville, FL, USA
E-mail: ismail@cnel.ufl.edu

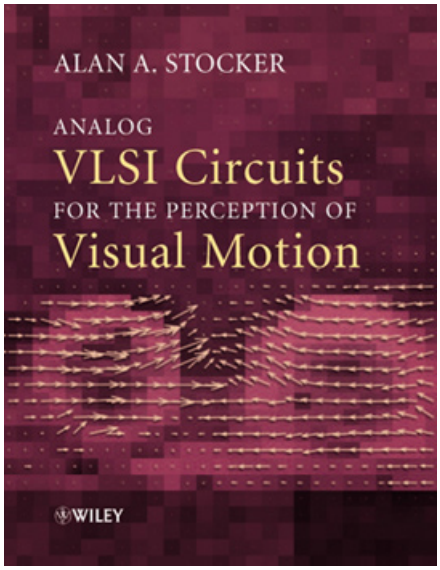
References:
1. I. Uysal, H. Sathyendra, and J. G. Harris, *A biologically plausible system approach for noise robust vowel recognition,* **IEEE Proc. of MWSCAS,** CD-ROM, 2006.
2. C. J. Sumner, E. A. Lopez-Poveda, L. P. O'Mard, and R. Meddis, *Adaptation in a revised inner-hair cell model,* **J. Acoust. Soc. Am. 113** (2), p. 893-901, 2003.
3. W. Maass, T. Natschlager, and H. Markram, *Real-time computing without stable states: A new framework for neural computation based on perturbations,* **Neural Computation 14** (11), pp. 2531-2560, 2002.

BOOK REVIEW

Analog VLSI Circuits for the Perception of Visual Motion

Alan A. Stocker, Wiley, March 2006

ISBN: 978-0-470-85491-4
Hardcover: 242 pages
US $130.00 / £70.00 / €105.00



Ever wondered why progress seems slow in building visually guided autonomous agents that perceive and intelligently interact with their environment? Well, one reason may be that our understanding of perception and the underlining computations involved is incomplete or just plain wrong. This new book by Alan Stocker provides *n* unconventional and fresh perspectives on how to understand perception and build simple artificial perceptual systems using analog VLSI (very large silicon integration) circuits. Focusing on the example of visual motion perception, it demonstrates how brain-style computation combined with CMOS (complimentary metal-oxide semiconductor technology can lead to efficient and robust 'neuromorphic' circuits to solve the hard optimization problems encountered in perception.

One key factor underlying the success of human visual perception lies in its use of constraint satisfaction. That is, the brain presumably applies mechanisms that combine the aspects of its visual input that cohere and segments out those aspects that do not. These mechanisms bootstrap globally coherent (optimal) solutions by rapidly satisfying local consistency constraints. Consistency depends on relative computations such as non-linear comparison, interpolation and error feedback, rather than absolute precision. And this style of computation is very suitable for implementation in analog VLSI circuits, as Dr. Stocker demonstrates.

What makes this book special is that it not only presents practical implementations of constraint satisfaction networks for visual motion perception, but it also demonstrates a series of useful and impressive aVLSI circuits for solving visual motion problems such as estimating 2D optical flow, motion segmentation, and motion selection. And these chips are useful for robotic applications. Their true strength lies, however, in their broad and principled theoretical foundations.

The book begins with some ecological considerations about why and how visual motion is perceived from changes in the visual input. It then goes on to illustrate the basic computational challenges, discusses possible solutions, and finally concludes in proposing a general computational architecture for visual motion perception. The key concept is that the perceptual process is an optimization problem of finding the visual motion estimate that is maximally consistent with the visual information and the system's expectations. Chapter three makes the connection to associative memory and Hopfield networks as examples of network architectures that compute optimal solutions. It demonstrates how simple problems (e.g. the winner-take-all operation) can be formulated as local constraints that together define the optimal solution. The chapter also shows how to derive appropriate network architectures that find it.

Chapter four then formulates optical-flow estimation as a constraint satisfaction problem, deriving the basic network architecture that is the basis for all further networks discussed in the book. It draws the connection between the formulated constraint solving problem and statistically optimal motion estimation as described with Bayesian frameworks, showing that prior information is essential in achieving a robust design. Furthermore, extensions of the basic network allow even more sophisticated processing such as motion segmentation or motion selection for which the network selects regions in its visual field that match a particular motion and size.

Chapters five to seven extensively deal with aVLSI implementations of the proposed network architectures, providing detailed schematics and measurements of the fabricated chips. The effects of the inevitable non-linearities and mismatch are discussed in detail, showing that clever analog designs can take advantage of nonlinearities to improve robustness and performance.

The book concludes with an interesting final chapter with a comparison to primate visual motion perception systems. It also presents data of head-to-head comparison between humans and the aVLSI chips performing the same perceptual tasks. The dynamics and steady-state behavior similarities are quite surprising, leading the author to conclude that both systems must optimize a similar set of constraints. The future will tell if this is true or not.

The broad approach of this book certainly reflects the background and the interests of the author. He is an expert aVLSI circuit designer, a computational modeler of the visual system, and a psychophysical experimentalist working on human motion perception. I highly recommend this book not only to those who are particularly interested in aVLSI visual motion circuits, but to anyone interested in the novel, neuromorphic, style of computation. The philosophy and methodology of the approach seem general enough and applicable to other perceptual tasks, such as depth perception and texture segmentation. Furthermore, the analog VLSI implementation of the presented computational networks becomes particularly attractive in light of recent technological developments in three- dimensional integrated circuits. Three-dimensional integration permits local vertical connections between different chips, physically stacked as a 'layer cake'. Recurrent analog networks can naturally be implemented as multi-layered parallel computational blocks of tremendous capabilities, without the need for sophisticated chip-to-chip protocols.

Ralph Etienne-Cummings
Department of Electrical and Computer Engineering
Johns Hopkins University
Baltimore, MD, USA
Email: retienne@jhu.edu
URL: http://etienne.ece.jhu.edu/

Artificial Brain, *continued from p. 1*

ficeMate will help users perform tasks such as scheduling, making telephone calls, data searching, and document preparation.

The auditory processor

The auditory module consists of feature-extraction, binaural and attention models, all inspired by the human auditory pathway. The feature extraction model is based on a cochlear filter bank, zero-crossing detector, and nonlinearity. The filter bank consists of many bandpass filters, of which center frequencies are distributed linearly in terms of the logarithmic scale. The zero-crossing time intervals are used to estimate robust frequency characteristics in noisy speeches. The logarithmic nonlinearity provides wide dynamic range and robustness to additive noise, while time-frequency masking may suppress weaker signals that are likely to be noise.

The binaural model estimates interaural time delay based on zero-crossing times for noise robustness. Also, the binaural processing algorithm has been extended to incorporate multiple sound sources and room acoustics with multipath reverberation. The convolutive ICA algorithm we developed successfully separates multiple speeches using linear or cochlear filterbanks.[1]

A simple but efficient top-down attention model has been developed with a multilayer Perceptron classifier for pattern recognition systems. In this top-down attention model, an attention cue may be generated either from the classified output or from an external source. The attended output class estimates an attended input pattern based on the top-down attention. It may be done by adjusting the attention gain coefficients for each input neuron using an error backpropagation algorithm. For unattended input features the attention gain may become very small, while those of attended features remains close to 1. Once a pattern is classified, attention may shift to find the remaining patterns.[2]

To provide the intensive computing power we developed a special chip for real-time applications. The system-on-a-chip consists of circuit blocks for analog to digital conversion, nonlinear speech-feature extraction, a programmable processor for the recognition system, and digital to analog conversion. Also, the extended binaural processing model has been implemented using field-programmable gate arrays and tested with a board with two microphones and five speakers (see Figure 2). The two microphones receive six audio signals, and the chip and board demonstrated great signal enhancement: the final signal-to-noise ratio was about 19dB, and the enhancement 18dB.[3]

The future

Intelligent machines will help humans as friends and family members in the early 21st century, and provide services for the prosperity of human beings. Intelligence to machines, and freedom to mankind!

Soo-Young Lee
Director, Brain Science Research Center
Dept. of BioSystems and
Dept. of Elec. Eng. and Comp. Sci.
Korea Advanced Inst. of Sci. and Tech.
Daejeon, Korea
E-mail: sylee@kaist.ac.kr

References

1. H.M. Park, S.H. Oh, and S.Y. Lee, *A filter bank approach to independent component analysis and its application to adaptive noise cancelling*, **Neurocomputing 55**, pp. 755-759, 2003.
2. B.T. Kim and S.Y. Lee, *Sequential recognition of superimposed patterns with top-down selective attention*, **Neurocomputing 58-60,** pp. 633-640, 2004.
3. C.M. Kim, H.M. Park, T. Kim, S.Y. Lee, and Y.K. Choi, *FPGA Implementation of ICA algorithm for blind signal separation and active noise canceling*, **IEEE Trans. on Neural Networks 14**, pp. 1038-1046, 2003.

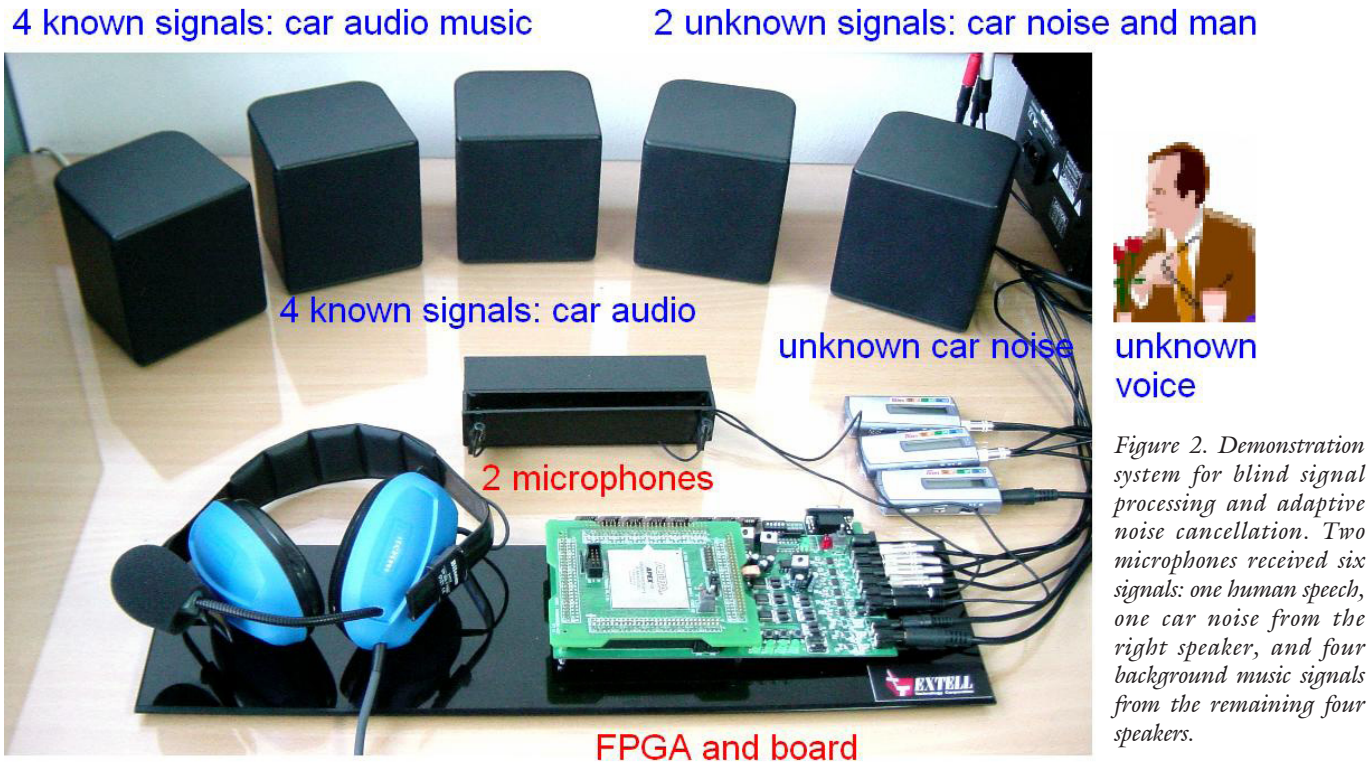


Figure 2. Demonstration system for blind signal processing and adaptive noise cancellation. Two microphones received six signals: one human speech, one car noise from the right speaker, and four background music signals from the remaining four speakers.