Artificial Brain and *OfficeMate^{TR}* based on Brain Information Processing Mechanism

Soo-Young Lee

Korea Advanced Institute of Science and Technology, Korea

Summary. The Korean Brain Neuroinformatics Research Program has dual goals, i.e., to understand the information processing mechanism in the brain and to develop intelligent machine based on the mechanism. The basic form of the intelligent machine is called Artificial Brain, which is capable of conducting essential human functions such as vision, auditory, inference, and emergent behavior. By the proactive learning from human and environments the Artificial Brain may develop oneself to become more sophisticated entity. The OfficeMate will be the first demonstration of these intelligent entities, and will help human workers at offices for scheduling, telephone reception, document preparation, etc. The research scopes for the Artificial Brain and OfficeMate are presented with some recent results.

1 Introduction

Although people had tried to understand the mechanism of brain for a long time, still only a few are understood. Even with the limited knowledge on biological brain artificial neural network researchers have come up with powerful information processing models and developed useful applications in the real world. Here we report an ambitious research program to develop human-like intelligent systems, called *'Artificial Brain'*, based on the brain information processing mechanism.

The Korean Brain Neuroinformatics Research Program got into the third phase in July 2004 for 4 years, which is regarded as the final phase of Korean brain national research program started in November 1998 for 10 years [1]. The program was initially sponsored by Ministry of Science and Technology, and is now sponsored by Ministry of Commerce, Industry, and Energy. It is a joint effort of researchers from many different disciplines including neuroscience, cognitive science, electrical engineering, and computer science, and currently about 35 PhDs and about 70 graduate students are involved in the program.

The Korean Brain Neuroinformatics Research Program has two goals, i.e., to understand information processing mechanisms in biological brains and to

Soo-Young Lee: *Artificial Brain and OfficeMate^{TR} based on Brain Information Processing Mechanism*, Studies in Computational Intelligence (SCI) **63**, 123–143 (2007)
www.springerlink.com

develop intelligent machines with human-like functions based on these mechanisms. In the third phase we are developing an integrated hardware and software platform for the brain-like intelligent systems, which combine all the technologies developed for the brain functions in the second phase. With two microphones, two cameras (or retina chips), and one speaker, the *Artificial Brain* looks like a human head, and has the functions of vision, auditory, cognition, and behavior. Also, with this platform, we plan to develop a testbed application, i.e., "artificial secretary" alias *OfficeMate*, which will reduce the working time of human secretary by a half.

In this chapter the goals and current status of the Korean Brain Neuroinformatics Research Program will be presented with some recent developments. The research goals and scopes are first described, and recent developments are presented latter.

2 Research Goals

To set up the research goals we incorporated two approaches, i.e., the bottom-up and top-down approaches, and set common goals for them. The bottom-up approach we incorporated is to extrapolate technology development trends and foresee future technology. The prediction of technology demands in future society has always been the main force of technology developments, and we regard it as the top-down approach.

Scientific progress has a tendency to be greatly influenced by unforeseeable breakthroughs, and the reliability of long-term prediction in scientific discovery is always questionable. However, it is still safe to say that recent developments in high performance brain imaging and signal processing equipment will greatly speed up understanding of brain architecture and information processing mechanisms. By reducing resolution both in time and space, it may be possible to record neuronal signals with sufficient accuracy for precise mathematical modeling. Although there still exists a big gap between the molecular neuroscience and system neuroscience, the existing gap between microscopic cellular models and macroscopic behavior models may eventually be bridged resulting in a unified brain information processing model.

Prediction in technology developments is regarded as more reliable. Especially, there exists a well-accepted theory, called "Moore's Law", in semiconductor industry. In 1965 Gordon Moore realized that each new chip contained roughly twice as many structures as its predecessor and each chip was released within 18–24 months of the previous chip. This trend is still remarkably accurate, and an increase rate of about 50 times is expected for the next 10 years. With more than 10 billion transistors in a single chip one may be able to implement a small part of human brain functions. More powerful systems may be built with multiple chips. Even in conventional computing architectures, the communication bottleneck between processors and memories will become more serious, and distributed computing and storage architectures will be pursued. Therefore, neuro-chip technology will fit into the main stream of computer and semiconductor industries.

Another interesting law, called “Machrone’s Law”, says that the machine you want always costs US$5,000. Actually it seems the cost is going down to US$1,000. People always wanted more powerful systems, and engineers had always come up with powerful systems with the same or lower price range. Therefore, the bottom-up prediction says that the enormous computing power will be available with affordable price in the future.

In human history the Industrial Revolution is regarded as the first big step to utilize machines for human welfare. With the help of powerful energy-conversion machines such as steam engines, the Industrial Revolution paved a road to overcome the physical limitation of humans and result in mass-production of consumer products. The second big step may be the Computer Revolution, which is based on electronic technology with accurate number crunching and mass data storage. Nowadays we can not imagine the world without the mass-production machines and computers.

What the future human society will be? People always wanted to resolve present difficulties and found ways to overcome the difficulties for the better life. The Machrone’s law may be a simple example. Although the Computer Revolution has provided better human life, it also creates problems, too. Computers are not yet sufficiently intelligent to work in a friendly way with humans. To make use of computers people must learn how to use it. In many cases it means learning tedious programming languages or memorizing the meaning of graphic user interface icons. Also, current computers do whatever they are programmed for, and do not have generalization and self-learning capabilities. Therefore, the programmers should take into account all possible cases for the specific application, and provide a solution for each case. Only a few people have these programming and logical-thinking abilities. To make computers useful to everybody, it is strongly recommended to make computers as human-friendly as possible. People shall be able to use computers as their friends or colleagues. Computers shall have human-like interfaces, self-learning capabilities, and self-esteem. The best way to accomplish this goal is to learn from the mother nature.

In Figure 1 information processing functions in brains are divided into 4 modules. A human has 5 sensors to receive information from environment, does some information processing based on these sensor signals, and provides motor controls. Among 5 sensors the vision and the auditory sensors provide the richest information, and complex information processing is performed. All the sensory information is integrated in the inference module, which provides learning, memory, and decision-making functions. The last module, Action Module, generates signals for required sensory motor controls. Although there may be many feedback pathways in biological brains, feed-forward signal pathways are mainly depicted here for simplicity.

Although the role of early vision systems is relatively well understood, we believe what we know about the brain is much less than what we do not

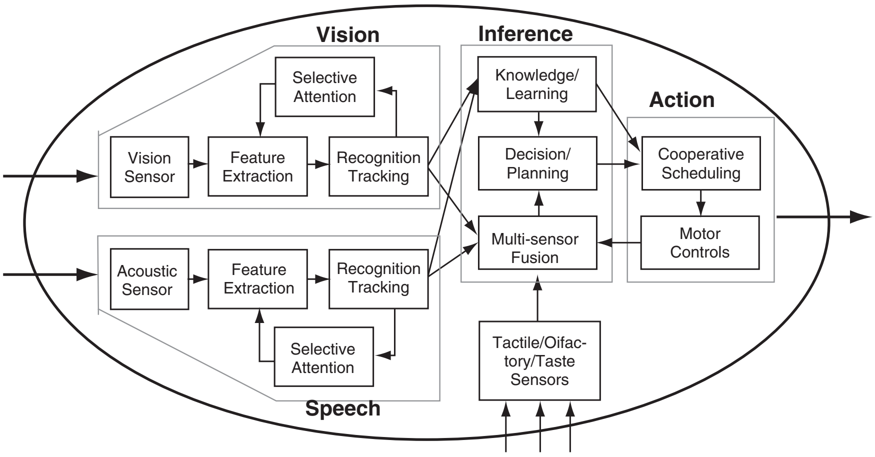

Fig. 1. Four functional modules in brain information processing systems. The *Artificial Brain* should also have 4 functional modules for vision, auditory, inference, and action systems

know. Compared to the vision and auditory modules the knowledge on the inference module is much more limited. However, even with a small hint from biological brains, we believe much more intelligent systems can be built. If neuroscientists concentrate on functions required to fill in the gaps of engineering functions, much faster progress may be achieved. Issues on invariant feature extraction, selective attention, adaptive dynamic ranges, sensory fusion, knowledge representation, generalization, self-learning, emotion, and cooperative behavior are only a few examples. Specific hardware implementations are also essential for the success. Therefore, a "system approach" is required to integrate efforts of researchers from many different disciplines for each module. Finally, the four modules need to be integrated as a single system, i.e., *Artificial Brain.*

The *Artificial Brain* may be trained to work for specific applications, and the *OfficeMate* is our choice of the application test-bed. Similar to office secretaries the *OfficeMate* will help users for office jobs such as scheduling, telephone calls, data search, and document preparation. The *OfficeMate* should be able to localize sound in normal office environment, rotate the head and cameras for visual attention and speech enhancement. Then it will segment and recognize the face. The lip reading will provide additional information for robust speech recognition in noisy environment, and both visual and audio features will be used for the recognition and representation of "machine emotion." The *OfficeMate* will use natural speech for communications with the human users, while electronic data communication may be used between *OfficeMates*. A demonstration version of the *Artificial Brain* hardware is shown in Figure 2.

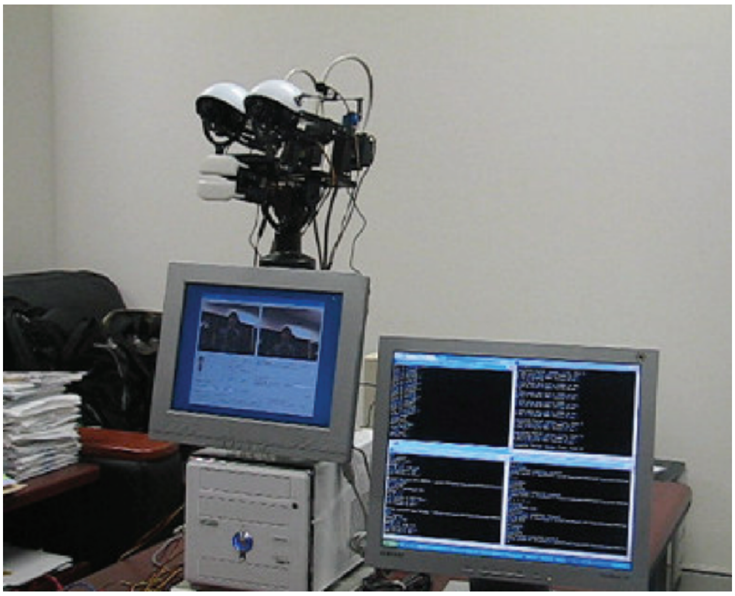

Fig. 2. *Artificial Brain* with two eyes, two ears, and one microphone. The lips are used for lip-sink, and 2 LCD displays are used for camera inputs and internal processor status

3 Research Scope

As shown in Figure 3 the *Artificial Brain* should have sensory modules for human like speech and visual capabilities, internal state module for the inference, and the output module for human-like behavioral control.

The sensory modules receive audio and video signals from the environment, and conduct feature extraction and recognition in the forward path. The backward path conducts top-down attention, which greatly improves the recognition performance of the real-world noisy speech and occluded patterns. The fusion of video and audio signals is also greatly influenced by this backward path.

The internal state module is largely responsible for intelligent functions and has a recurrent architecture. The recurrent architecture is required to model human-like emotion and self-esteem. Also, the user adaptation and proactive learning are performed at this internal state module.

The output module generates human-like behavior with speech synthesizer and facial representation controller. Also, it provides computer-based services for *OfficeMate* applications.

In the Korean Brain Research Program we are trying to develop detail mathematical models for the Artificial Brain. In the mathematical model the internal state value $\mathbf{H}[n+1]$ is defined as a function of sensory inputs, previous internal states, and system outputs, i.e.,

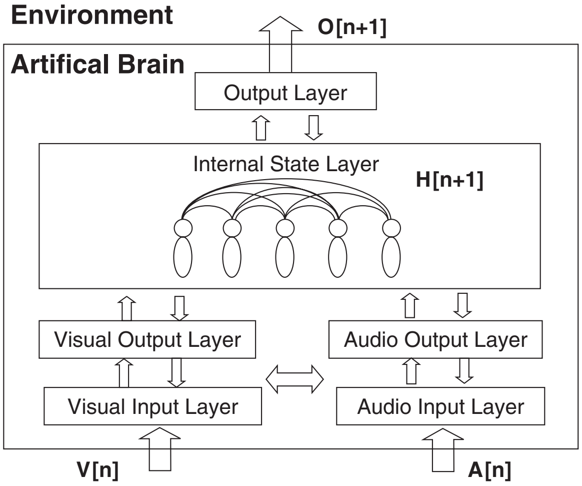

Fig. 3. Block diagram of *Artificial Brain* and its interaction with the environment

$$\boldsymbol{H}[n+1] = \boldsymbol{f}(\boldsymbol{V}[n], \boldsymbol{A}[n], \boldsymbol{H}[n], \boldsymbol{O}[n]), \qquad (1)$$

where $\mathbf{V}[n], \mathbf{A}[n], \mathbf{H}[n]$, and $\mathbf{O}[n]$ denote video inputs, audio inputs, internal state values, and outputs at time n, respectively, and $f(\cdot)$ is a nonlinear function. The output is defined as a function of the internal states and sensory inputs, i.e.,

$$\boldsymbol{O}[n+1] = \boldsymbol{g}(\boldsymbol{H}[n+1],\ \boldsymbol{V}[n], \boldsymbol{A}[n]), \qquad (2)$$

where $\boldsymbol{g}(\cdot)$ is a nonlinear function. It is also worth noting that the sensory inputs are functions of both the system outputs and environment states as

$$\boldsymbol{V}[n] = \boldsymbol{p}(\boldsymbol{O}[n],\ \boldsymbol{E}[n]),\ \boldsymbol{A}[n] = \boldsymbol{q}(\boldsymbol{O}[n], \boldsymbol{E}[n]), \qquad (3)$$

where $\mathbf{E}[n]$ is the environment state value and $\boldsymbol{p}(\cdot)$ and $\boldsymbol{q}(\cdot)$ are nonlinear functions.

Although the basic technologies had been developed for the visual and audio perception during the last 8 years, the most challenging part is the development of the "Machine Ego" with human-like flexibility, self-learning performance, and emotional complexity. It will also have user-modeling capabilities for practical user interfaces. We believe the *Artificial Brain* should have active learning capability, i.e., the ability to ask "right" questions interacting with people. To ask right questions the *Artificial Brain* should be able to monitor itself and pinpoint what it may need to improve. Based on this observation we would like to develop a mathematical model of the Machine Ego, which is the most important component of the *Artificial Brain*. Research scopes for the four modules are summarized as follows.

3.1 Auditory Module

The research activities on the auditory module are based on the simplified diagram of the human auditory central nervous system. Detail functions currently under modeling are summarized in Figure 4. The object path, or "what" path, includes nonlinear feature extraction, time-frequency masking, and complex feature formation from cochlea to auditory cortex. These are the basic components of speech feature extraction for speech recognition. The spatial path, or "where" path, consists of sound localization and noise reduction with binaural processing. The attention path includes both bottom-up (BU) and top-down (TD) attention. However, all of these components are coupled together. Especially, the combined efforts of both BU and TD attention control the object and spatial signal paths.

The nonlinear feature extraction model is based on cochlear filter bank and logarithmic nonlinearity. The cochlear filter bank consists of many bandpass filters, of which center frequencies are distributed linearly in logarithmic scale. The quality factor Q, i.e., ratio of center frequency to bandwidth, of bandpass filters is quite low, and there are overlaps in frequency characteristics. The logarithmic nonlinearity provides wide dynamic range and robustness to additive noise. Time-frequency masking is a psychoacoustic phenomenon,

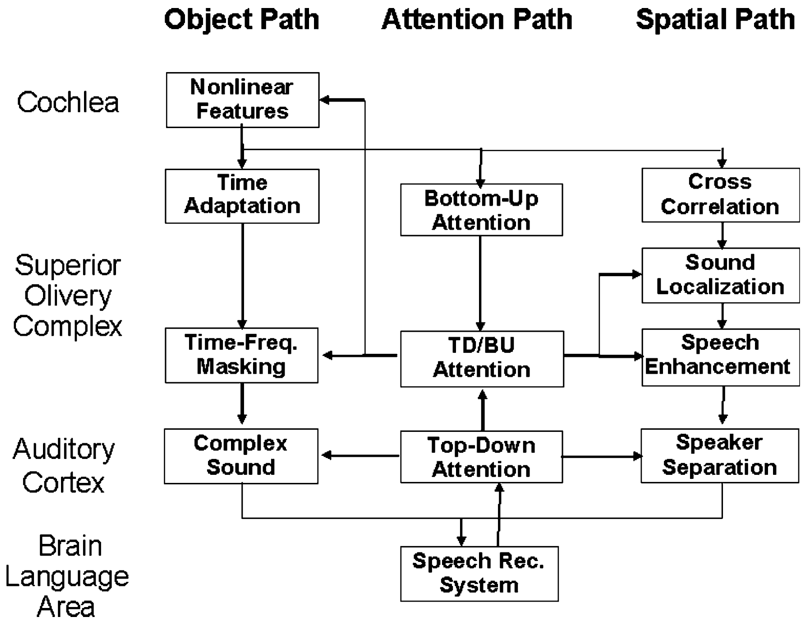

Fig. 4. Block diagram of auditory pathway model. The object path and spatial path deal with speech feature extraction and sound localization, respectively, and the attention path controls the other two paths for robust speech recognition

where a stronger signal suppresses weaker signals in nearby time and frequency domains.

For the binaural processing at the spatial path conventional models estimate interaural time delay, i.e., time-delay between signals from left and right ears, based on cross-correlation, and utilize the time-delay for sound localization and noise reduction. Interaural intensity difference is also utilized for advanced models. However, these models assume only direct sound paths from a sound source to two ears, which is not valid for many real-world environments with multipath reverberation and multiple sound sources (e.g., speech inside an automobile with external road and wind noise, and reverberation of speech mixed with music from the audio system). Therefore, it is required to incorporate deconvolution and separation algorithms in the binaural processing.

For the attention path, a model is being developed to combine both the bottom-up (BU) and top-down (TD) attention mechanisms. The BU attention usually results from strong sound intensity and/or rapid intensity changes in time, and is closely related to the time-frequency masking. However, TD attention comes from familiarity and importance of the sound, and relies on existing knowledge of each person. For example, a specific word or a person's voice may trigger TD attention for relevant people only. Therefore, TD attention originates from the higher-level brain areas that may be modeled in a speech recognition system.

3.2 Vision Module

The vision module also consists of submodules for feature extraction in the object path, stereo vision in the spatial path, and image recognition in the attention path. Also, it is closely coupled to the action module for the active vision and facial representation of emotion.

The object path starts from the bottom-up saliency map [2] to identify the area of interests, and pattern recognition with top-down attention is performed only at those areas. The saliency map consists of colors and orientation edges with several different scales. The recognition submodule will visit each area with high saliency one by one, and classify the images. In the first version of *Artificial Brain* the vision module mainly identifies the facial areas from background images, and recognizes the name and emotional status of the person. Similar to the auditory module the top-down attention greatly improves the recognition performance of occluded or confusing patterns in complex backgrounds.

An important research topic for this module is the color constancy with different illumination conditions.

In future lip-reading will be added for robust recognition in very noisy environment. Since the human perception of motion goes through two different pathways, i.e., the lateral fusiform gyrus for the invariant aspects and the superior temporal sulcus for the changeable aspects of faces [3], the dynamic video features may be different from static image features, and efficient unsupervised learning algorithm should be developed to extract the dynamic features.

3.3 Inference Module

The inference module performs knowledge acquisition, emotional transition, and user adaptation. Applications of inference functions for *OfficeMates* are also integrated in this module.

The knowledge acquisition should be autonomous and proactive. For the autonomous learning it should be able to learn without intervention of users. For example, if a textbook on medicine is provided, the *Artificial Brain* should be able to learn the knowledge of medical doctors. To accomplish this goal we develop unsupervised learning algorithms to extract the basic components of knowledge from the text. A hierarchical architecture may be adopted to build complex knowledge systems from these basic components. The proactive learning then improves the knowledge by asking proper questions. The module estimates what need to be learnt more, phrases proper questions, figures out appropriate person to ask, and incorporates the answers into its knowledge system.

Even with the proactive learning the inference module may experience new circumstances that it has never been exposed to before in the real world applications. Therefore, another important characteristic of the learning system is the generalization capability, which may be obtained by additional constraints on the cost function during learning [4].

The emotional transition is one important characteristic of human-like behavior. To incorporate the emotional transitions we use recurrent neural networks in the inference module, and one hidden neuron is assigned to each independent emotional axis. The transition may be triggered by sensory perception and its own actions to the environment. However, in the future the emotion assignment and transition will be learnt autonomously, and the efficient learning algorithm of this emotional network still need be investigated. If successful, it may lead us to the more complex topics of consciousness and self esteem.

The user adaptation has many different levels, and the simplest level may be implemented by adjusting some parameters of the inference system. However, we plan to implement the user adaptation as the training of another inference system for the user. In this framework both the *Artificial Brain* and users share the same inference architecture, and the two inference modules are learnt simultaneously.

The applications of the inference module include language understanding, meeting scheduling, and document preparation. Actually the language understanding is the fundamental function for efficient man-machine interface. Also, the extraction of emotional components from speech and texts is conducted during this process.

The *Artificial Brain* need to be trained for specific applications of the *OfficeMates*. We focus on two jobs of office secretaries, i.e., meeting scheduling and document preparation. Of course we do not expect perfect performance at the early stage, but hope to save time of human secretaries by a half.

3.4 Action Module

The action module consists of speech synthesizer and facial representation controller. Both are expected to have capabilities of emotional representation, which is very important for the natural interaction between the *Artificial Brain* and its users. The speech synthesizer is based on commercial TTS (Text-To-Speech) software, and we are just adding capability of emotional speech expressions. The emotional facial representation has been analyzed, and the robot head of the *Artificial Brain* is capable of representing simple emotions.

Another important function of the action module is the communication with other office equipments such as telephone, computer, fax machine, copier, etc. Although it does not require intelligence, it is needed to work as *OfficeMates*.

4 Research Results

In this section some research results are reported mainly for the auditory and vision modules. The inference and action modules are still at the early stage of research.

4.1 Self-Organized Speech Feature

The nonlinear feature extraction in auditory pathway is based on cochlear filter bank and logarithmic nonlinearity. The cochlear filter bank consists of many bandpass filters, of which center frequencies are distributed linearly in logarithmic scale. Based on the information-theoretic sparse coding principle we present the frequency-selective responses at the cochlea and complex time-frequency responses at the auditory cortex.

At cochlea we assume that speech signal is a linear combination of the independent basis features, and find these basis features by unsupervised learning from the observed speech. The Independent Component Analysis (ICA) minimizes the mutual information and extracts the statistically independent features [5]. For speech signals we assume the Laplacian distribution, of which sparsity was supported by an experiment on the dynamic functional connectivity in auditory cortex [6].

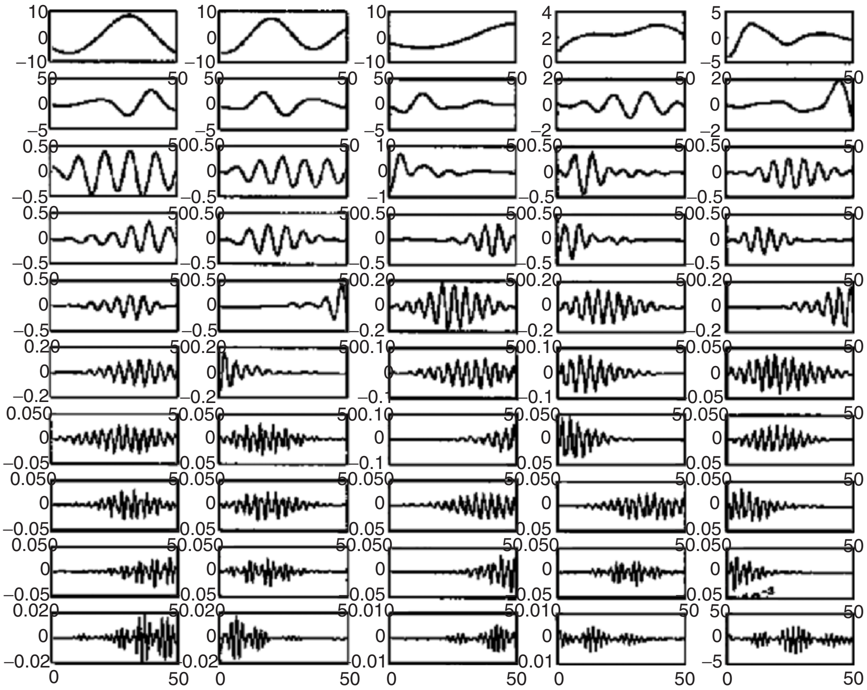

Fig. 5. Fifty simple speech features extracted by independent component analysis from raw speech signals

The training data consist of 60 sentences from six speakers in the TIMIT continuous speech corpus (http://www.ldc.upenn.edu/Catalog/docs/LDC93S2/timit.html), and speech segments composed of 50 samples, i.e., 10 ms time interval at 16 kHz sampling rates, are randomly generated.

As shown in Figure 5, the obtained 50 basis feature vectors are localized in both time and frequency [7]. Average normalized kurtosis of the extracted features is 60.3, which shows very high sparseness. By applying the topology-preserving ICA [8], the basis features are extracted in the order of the center frequency [9].

After the frequency-selective filtering at the cochlea, more complex auditory features are extracted at the latter part of the human auditory pathway, i.e., inferior colliculus and auditory cortex. This complex features may also be understood as the information-theoretic sparse coding. Here we model the earlier part of the human auditory pathway as a simple mel-scaled cochlear filterbank and the logarithmic compression. The time-frequency representation of the speech signal is estimated at each time frame with 10 msec intervals, and the ICA algorithm is applied to this two-dimensional data. The 23 mel-scaled filters and 5 time frames are selected with the local feature dimension of 115, which is reduced to 81 using the principal component analysis (PCA). Topology-preserving self-organization is also incorporated [8].

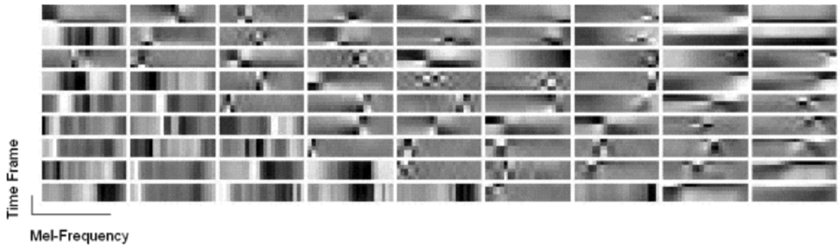

Fig. 6. Eighty one complex speech features extracted by independent component analysis from time-frequency spectrum

As shown in Figure 6, the resulting complex speech features show many aspects of the features extracted at the auditory cortex [10]. At the lower left side of the map, vertical lines represent frequency-maintaining components with complex harmonics structures. The horizontal lines at the upper right side of the map represent on-set and off-set components. In the center of the map, there exist frequency-modulation components such as frequency-rising and frequency-falling components. In fact, there exist neurons responding to these three basic sound components in the human auditory pathways, i.e., the steady complex harmonics, on/off-sets, and frequency modulation. Many auditory cortical areas are tonotopically organized, and are specialized to specific sound features [11].

4.2 Time-Frequency Masking

Another important characteristic of the signal processing in the human auditory pathway is the time-frequency masking, which had been successfully modeled and applied to the noise-robust speech recognition [12]. Time-frequency masking is a psychoacoustic phenomenon, where the stronger signal suppresses the weaker signals in nearby time and frequency domains [13]. It also helps to increase frequency selectivity with overlapping filters.

As shown in Figure 7, the frequency masking is modeled by the lateral inhibition in frequency domain, and incorporated at the output of the Mel-scale filterbank. The time masking is also implemented as lateral inhibition, but only the forward (progressive) time masking is incorporated.

The developed time-frequency masking model is applied to the isolated word recognition task. Frequency masking reduces the misclassification rates greatly, and the temporal masking reduces the error rate even further [12].

4.3 Binaural Speech Separation and Enhancement

For the binaural processing the usual model estimates interaural time delay based on cross-correlation, and utilizes the time-delay for sound localization and noise reduction. Interaural intensity difference is also utilized for advanced

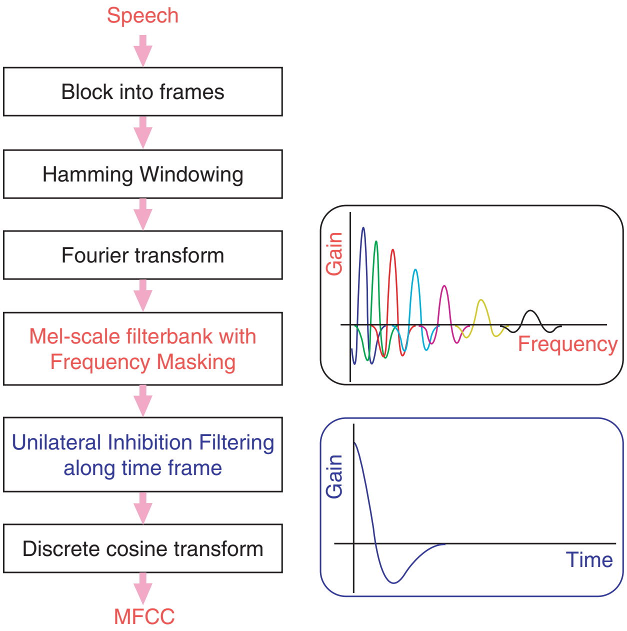
Speech
Block into frames
Hamming Windowing
Fourier transform
Mel-scale filterbank with Frequency Masking
Unilateral Inhibition Filtering along time frame
Discrete cosine transform
MFCC
Gain
Frequency
Gain
Time

Fig. 7. Block diagram of time-frequency masking and their lateral interconnections

models [13]. However, these models assume only direct sound paths from one sound source to two ears, which is not true for many real-world environments with multipath reverberation and multiple sound sources. Therefore, it is required to incorporate deconvolution and separation algorithms, and an extended binaural processing model has been developed based on information theory. We have extended the convolutive ICA algorithm [14] to multiple filterbanks [15] and further extended the cochlea filterbank.

As shown in Figure 8, the signals from the left ear and the right ear first go through the same filterbank, and the outputs of each filter are de-mixed by separate ICA networks. Then, the clean signals are recovered through inverse filterbanks. If two signal sources exist, each signal can be recovered. If only one signal source exists, the signal and a noise will be recovered.

In ref. [15] the ICA-based binaural signal separation with uniform filterbank results in much higher final SIR than the fullband time-domain approach and the frequency-domain approach. The poor performance of the frequency-domain approach comes from the boundary effects of the frame-based short-time Fourier transform as well as the permutation problem of the ICA algorithm. Although the permutation problems still needs to be solved, compared to the standard time-domain approach without the filterbank, the filterbank approach converges much faster giving better SNR. Basically the filterbank approach divides the complex problem into many easier problems. Due to the decimation at each filter the computational complexity is also

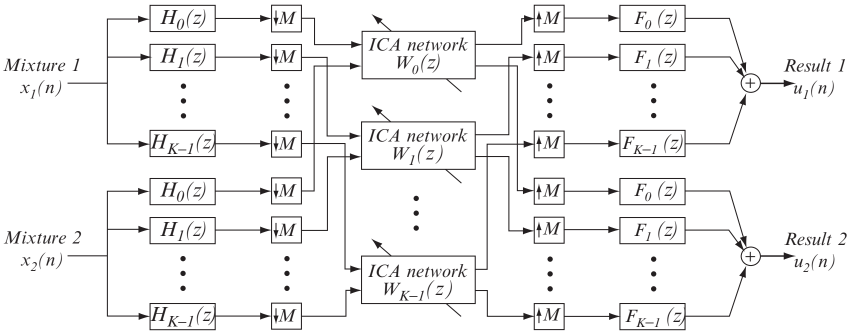

Fig. 8. Binaural processing model with filterbank and independent component analysis. Two microphone signals first go through bandpass filterbank, and separate ICA network is applied to each filtered signals. The filtered signals may be combined by inverse filters

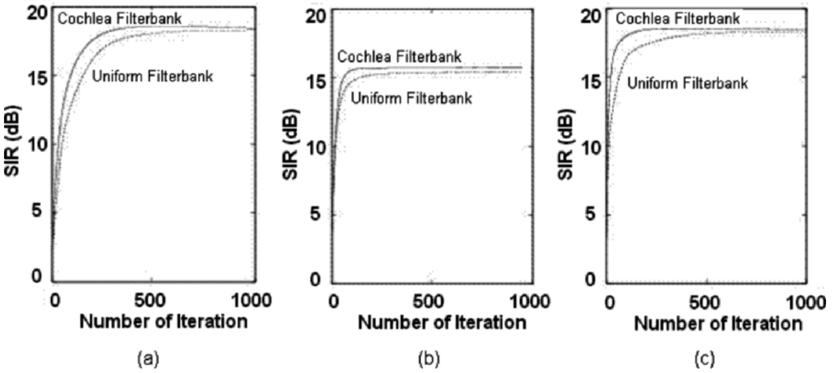

Fig. 9. Performance of ICA-based binaural signal separation methods from convolutive mixtures. The ICA with cochlea filterbank converges much faster than uniform filterbank approach. (a) Speech and music; (b) two speeches; (c) speech and car noise

reduced. Also, it is more biologically plausible. As shown in Figure 9, the utilization of cochlea filterbank greatly improves the convergence.

The de-mixing matrices include information on the relative positions of the sources from the microphones, and the sound localization is also achievable from the de-mixing coefficients. The filterbank approach is quite advantageous for the accurate estimation of the sound direction, especially for noisy multi-source cases. Also, the estimated sound direction may also be utilized to solve the permutation problem [16].

4.4 Top-Down Selective Attention

In the cognitive science literature two different processes are presented with the word “selective attention”, i.e., the bottom-up (BU) and top-down (TD) attention mechanisms. The BU attention usually incurs from strong sound intensity and/or fast intensity changes in time. However, the TD attention comes from familiarity and perceptual importance of the sound, and relies on existing knowledge of each person. For example, a specific word or a person's voice may trigger TD attention for relevant people only.

The TD attention originates from the higher brain areas, which may be modeled as a speech recognition system. A simple but efficient TD attention model has been developed with a multilayer perceptron classifier for the pattern and speech recognition systems [17][18]. As shown in Figure 10, the sensory input pattern is fed to a multi-layer Perceptron (MLP), which generates a classified output. Then, 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 by error backpropagation. 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, the attention shifting may occurs to find the remaining patterns. In this case the attention gain coefficients of highly-attended features may be set to 0, while the other may be adapted.

The main difficulty of this top-down expectation comes from the basic nature of the pattern classification. For pattern classification problems many input patterns may belong to the same output class, and the reverse is not unique. However, for many practical applications, one only needs to find the closest input pattern to the attended class, and the gradient-descent algorithm does just that.

Figure 11 shows examples of selective attention and attention switching algorithm in action for confusing patterns [19] and overlapped numerals [17].

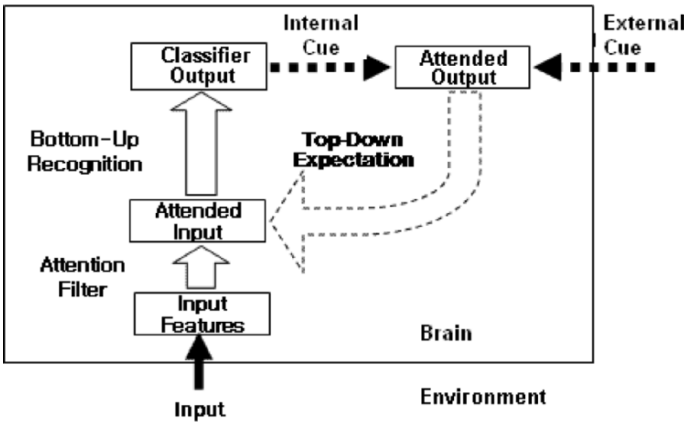

Fig. 10. Block diagram of top-down attention mechanism. The top-down expectation is estimated from the attended output class by the multi-layer perceptron classifier, which mimics the previous knowledge on the words and sounds

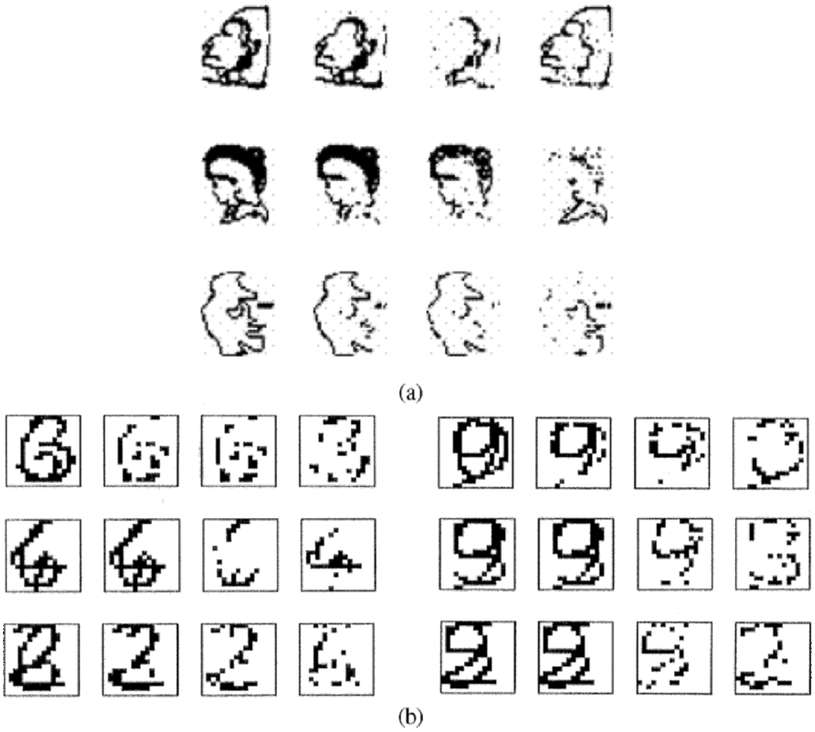

Fig. 11. Examples of selective attention and attention switching. The four images in each row show from the left the test input, attended input, attention-switched input, and the second-round input, respectively. (a) Results for 3 confusing images, i.e. Eskimo and the facial side view, lady face and old man face, and trumpet player and facial front view; (b) results from overlapped 10 numeric characters

The four images on the horizontal sequences show results on one test. The first image shows the confusing or overlapped test pattern. The second image shows the attended input for the first round classification. The third image shows the masking pattern for attention switching. The fourth image shows the residual input pattern for the second round classification. Figure 11 clearly shows that selective attention and attention switching are performed effectively, and the remaining input patterns for the second round classification are quite visible. The top-down attention algorithm recognized much better than standard MLP classifier, and the attention shifting successfully recognized two superimposed patterns in sequence. It also achieved much better recognition rates for speech recognition applications in real-world noisy environment [18].

We also combined the ICA-based blind signal separation and top-down attention algorithms [20]. The ICA algorithm assumes that the source signals are statistically independent, which is not true for many real-world speech signals. Therefore, the ICA-based binaural signal separation algorithm results in non-exact source signals. By incorporating attention layer at the output of

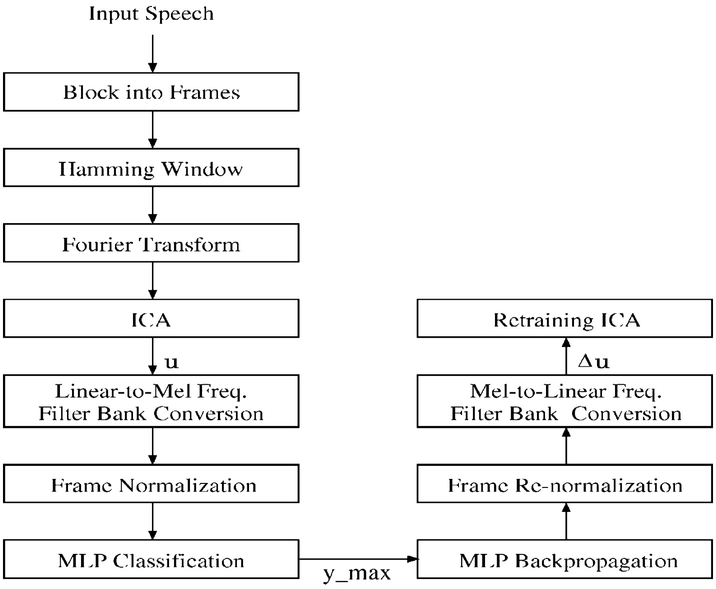

Fig. 12. Block diagram of ICA-based signal separation with deviation correction from top-down attention. It may be understood as a BU-TD combined approach, in which the ICA network serves for the bottom-up attention

the ICA network, this deviation may be compensated for the reference signal provided by the top-down attention. For speech recognition tasks the Mel-Frequency Cepstral Coefficient (MFCC) feature is the popular choice, and the backward evaluation becomes complicated. However, as shown in Figure 12, it is still applicable. Basically one may consider the calculation steps of the MFCC as another layer of a nonlinear neural network, and apply the error backpropagation with the specific network architecture.

4.5 Dynamic Features for Lip-reading

In previous studies the lip-motion features are extracted from single-frame images and the sequential nature of the motion video is not utilized. However, it is commonly understood that the human perception of static images and motion go through different pathways. The features of motion video may be different from the features for the face recognition, and requires more representation from consecutive multiple frames.

Figure 13 shows the dynamic features extracted by 3 decomposition techniques, i.e., Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF), and Independent Component Analysis (ICA), from multi-frame lip videos [21]. While the PCA results in global features, the ICA results in local features with high sparsity. The sparsity of the NMF-based features resides between those of the PCA and ICA-based features. The

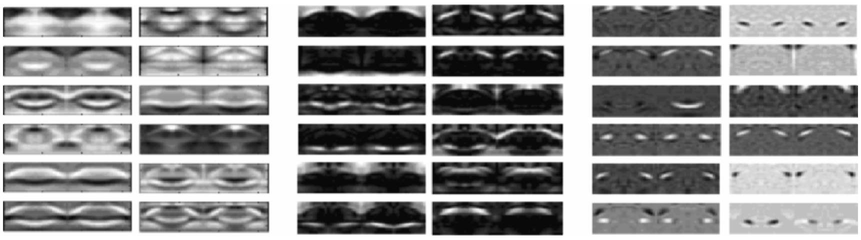

Fig. 13. Extracted lip motion features by PCA (left figures), NMF (center figures), and ICA (right figures) algorithms. Only features from 2-frames are shown for simplicity

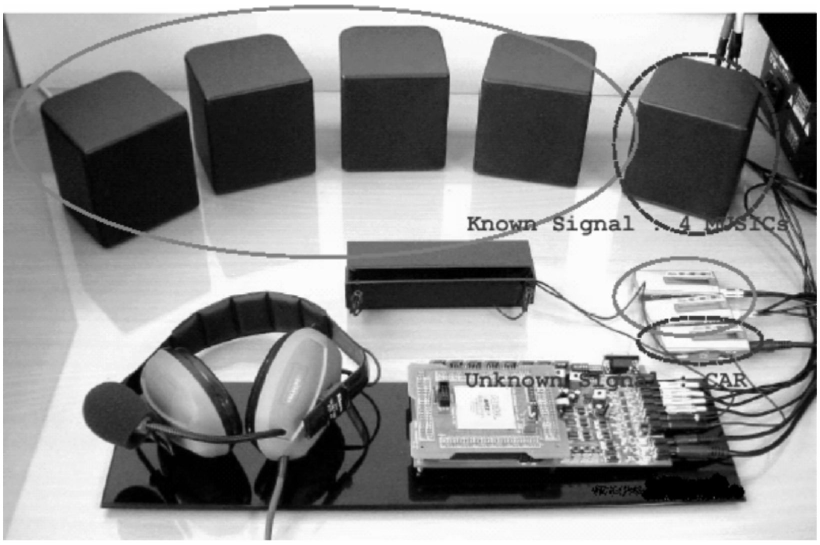

Fig. 14. Demonstration system for the blind signal processing and adaptive noise canceling. Two microphones received 6 signals, i.e., one human speech, one car noise from the right speaker, and 4 background music signals from the remaining 4 speakers

probability density functions and kurtosis of these features are almost independent on the number of the consecutive frames, and the multiple-frame features require less number of coefficients to represent video clips than the single-frame static features. It was also found that the ICA-based features result in the best recognition performance for the lip-reading.

4.6 Hardware Implementations

Many auditory models require intensive computing, and special hardware has been developed for real-time applications. A speech recognition chip had been developed as a System-On-Chip, which consists of circuit blocks for AD conversion, nonlinear speech feature extraction, programmable processor for recognition system, and DA conversion. Also, the extended binaural processing model has been implemented in FPGAs [22].

The developed FPGA-chip was tested with a board with two microphones and 5 speakers. Four of these speakers mimic car audio signals, of which original waveforms are available from electric line jacks. The other speaker generates car noise signal. Also, there is another human speaker. Therefore, the two microphones receive 6 audio signals as shown in the upper part of Figure 14. The developed chip and board demonstrated great signal enhancement, and result in about 19 dB final SNR or 18 dB enhancements. The performance of the FPGA-chip is tested for speech recognition tasks, and the achieved recognition rates are almost the same as those of a clean speech.

5 The Future

The intelligent machines will help human as friends and family members in the early 21st century, and provide services for the prosperity of human beings. In 2020 each family will have at-least one intelligent machine to help their household jobs. At offices intelligent machines, such as the *OfficeMates*, will help human to work efficiently for the organizations. We expect the number of working people may be reduced by a half with the help of *OfficeMates*, and the other half may work on more intelligent jobs. Or, they may just relax and enjoy their freedom.

Intelligence to machines, and freedom to mankind!

Acknowledgment

This research has been supported by the Brain Neuroinformatics Research Program by Korean Ministry of Commerce, Industry, and Commerce. The author also would like to represent his sincere thanks to his former and current students who had contributed to the researches.

References

[1] Lee, S.Y.: Korean Brain Neuroinformatics Research Program: The 3rd Phase. International Joint Conference on Neural Networks, Budapest, Hungary (2004).

[2] Itti L., Koch, C.: Computational model of visual attention. Nature Reviews Neuroscience 2 (2001) 194–203.

[3] Haxby, J.V., Hoffman, E.A., Gobbini, M.I.: The distributed human neural system for face perception. Trends in Cognitive Sciences 4 (2000) 223–233.

[4] Jeong, S.Y., Lee, S.Y.: Adaptive learning algorithm to incorporate additional functional constraints into neural networks. Neurocomputing 35 (2000) 73–90.

[5] Olshausen, B., Field, D.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381 (1996) 607–609.

[6] Clement, R.S., Witte, R.S., Rousche, P.J., Kipke, D.R.: Functional connectivity in auditory cortex using chronic, multichannel unit recordings. Neurocomputing 26 (1999) 347–354.

[7] Lee, J.H., Lee, T.W., Jung, H.Y., Lee, S.Y.: On the Efficient Speech Feature Extraction Based on Independent Component Analysis. Neural Processing Letters 15 (2002) 235–245.

[8] Hyvarinen, A., Hoyer, P.O., Inki, M.: Topographic independent component analysis. Neural Computation 13 (2001) 1527–1558.

[9] Jeon, H.B., Lee, J.H., Lee, S.Y.: On the center-frequency ordered speech feature extraction based on independent component analysis. International Conference on Neural Information Processing, Shanghai, China (2001) 1199–1203.

[10] Kim, T., Lee, S.Y.: Learning self-organized topology-preserving complex speech features at primary auditory cortex. Neurocomputing 65-66 (2005) 793–800.

[11] Eggermont, J.J.: Between sound and perception: reviewing the search for a neural code. Hearing Research 157 (2001) 1–42.

[12] Park, K.Y., Lee, S.Y.: An engineering model of the masking for the noise-robust speech recognition. Neurocomputing 52-54 (2003) 615–620.

[13] Yost, W.A.: Fundamentals of hearing – An introduction. Academic Press (2000).

[14] Torkkola, T.: Blind separation of convolved sources based on information maximization. In Proc. IEEE Workshop on Neural Networks for Signal Processing, Kyoto (1996) 423–432.

[15] Park, H.M., Jeong, H.Y., Lee, T.W., Lee, S.Y.: Subband-based blind signal separation for noisy speech recognition. Electronics Letters 35 (1999) 2011–2012.

[16] Dhir, C.S., Park, H.M., Lee, S.Y.: Permutation Correction of Filter Bank ICA Using Static Channel Characteristics. Proc. International Conf. Neural Information Processing, Calcutta, India (2004) 1076–1081.

[17] Lee, S.Y., Mozer, M.C.: Robust Recognition of Noisy and Superimposed Patterns via Selective Attention. Neural Information Processing Systems 12 (1999) MIT Press 31–37.

[18] Park, K.Y., and Lee, S.Y.: Out-of-Vocabulary Rejection based on Selective Attention Model. Neural Processing Letters 12 (2000) 41–48.

[19] Kim, B.T., and Lee, S.Y.: Sequential Recognition of Superimposed Patterns with Top-Down Selective Attention. Neurocomputing 58-60 (2004) 633–640.

[20] Bae, U.M., Park, H.M., Lee, S.Y.: Top-Down Attention to Complement Independent Component Analysis for Blind Signal Separation. Neurocomputing 49 (2002) 315–327.

[21] Lee, M., and Lee, S.Y.: Unsupervised Extraction of Multi-Frame Features for Lip-Reading. Neural Information Processing – Letters and Reviews 10 (2006) 97–104.

[22] Kim, C.M., Park, H.M., Kim, T., Lee, S.Y., Choi, Y.K.: FPGA Implementation of ICA Algorithm for Blind Signal Separation and Active Noise Canceling. IEEE Transactions on Neural Networks 14 (2003) 1038–1046.