Chapter 5
A Framework for Intelligent "Conscious" Machines Utilising Fuzzy Neural Networks and Spatial-Temporal Maps and a Case Study of Multilingual Speech Recognition

Nikola Kasabov

Department of Information Science
University of Otago, P.O Box 56, Dunedin, New Zealand
Phone: +64 3 479 8319, fax: +64 3 479 8311
nkasabov@otago.ac.nz

Abstract. This chapter contains a discussion material and preliminary experimental results of a new approach to building intelligent conscious machines (ICM) and its application to multilingual spoken recognition systems. ICM can analyse their behaviour and subsequently adapt and improve their structure and functionality during operation, can evaluate their ability of problem solving in terms of what they "can do" and what they cannot. These systems consist of many modules interacting during operation and organised in several hierarchical levels aggregated into two main ones, a low, sub-conscious level, and a higher, conscious level. A framework for intelligent conscious machines is proposed and a partial realisation is presented which makes use of fuzzy neural networks and spatial-temporal maps. The framework is applicable to recognising patterns from time-series at different time scales, with numerous applications. A particular case study of spoken language recognition is presented along with some preliminary experimental results of a system realisation. The approach, introduced in the chapter, is extended to multi-lingual spoken recognition systems. This has been inspired by new biological evidence about the activity of the human brain in multi-lingual subjects.
Key words: intelligent information systems, multi-lingual speech recognition, cognitive engineering, fuzzy neural networks, conscious machines.

1. Introduction: Intelligent Information Systems and Intelligent Conscious Machines

This section specifies the scope of AI, cognitive engineering and brain-like computing as used in the chapter, and also defines the notions of intelligent information systems (IIS) and intelligent conscious machines (ICM).

Intelligence is usually associated with such characteristics as:

- ability to communicate ideas and thoughts in speech and language;

- pattern recognition, e.g. speech patterns, images, time series events;
- learning from structured and unstructured experience and successful generalisation;
- dynamic adaptation to new situations;
- reasoning and decision making based on uncertainty;
- creativity, i.e. creating something which is missing at present; e.g. plans.

Intelligent information systems (IIS) have some or all of the characteristics above in addition to having large memories and fast "number crunching" abilities. The combination of the computing power of the traditional computers with computational intelligence makes the IIS very powerful a means for information processing. Adding "consciousness" to an IIS would enable these machines to be aware of what they are in the current operating environment, how they relate to other objects, what they can do and what they can not, who might do what they can not attempt, etc. (Arbib, 87, 95; Aleksander, 97). We shall call such machines *intelligent conscious machines* (ICM).

Developing methods and tools for IIS, including ICM, based on cognitive principles of the brain, is the area of cognitive engineering. Some of the cognitive features of an IIS may be achieved by using connectionist techniques and principles adopted from the physical organisation of the brain, which is the area of the brain-like computing (Arbib, 95). Other cognitive features will be implemented by using other (non-brain-like) techniques such as symbolic AI, fuzzy logic, etc. (Zadeh, 65,84; Kasabov, 96).

IIS at present are usually realised as hybrid AI systems i.e. they make use of several AI paradigms in one system. Hybrid systems, consisting of a low, sub-conscious level, and a higher, conscious level, have been suggested by several authors (Kasabov, 90; Handelman, Lane and Gefland, 90; Hendler and Dickens, 91). A two-level hierarchical framework is suggested in (Kasabov, 90, 96) and shown in fig.1.1. The block diagram on fig.1.1 is simple but has a sophisticated functionality as explained here. ‘The first (low) level communicates with the environment, recognises images and more complex situations, learns new knowledge in a stimulus-reaction way, etc., but the final solution is communicated at a higher level, which performs "deliberate thinking", planning and symbol processing. The low level is fast, flexible, adaptable, subconscious. The high level is slow, serial, conscious. The low level operates mainly with numbers, values and connections. The high level is mainly symbolic. It operates with objects, relations, concepts, rules, premises, hypotheses, plans, strategies and classes, etc. Both levels communicate until the final solution is reached. Each of the two main levels of the general two-level model, can consist of more sub-levels.

The framework as described above was realised in several hybrid connectionist rule-based systems: COPE (Kasabov, 93), FuzzyCOPE (Kasabov, 95, 96; Kasabov, Kim, et al, 97). Other hybrid systems which combine symbolic AI, fuzzy logic, neural networks and possibly some other AI paradigms were reported in (Takagi, 90; Yamakawa, 93; Hashiyama, Furuhashi and Uchikawa, 92).

Fig. 1.1. A general two-level hierarchical model of an IIS (Kasabov,90, 96)

A significant limitation of these systems is the way rules are interpreted, that is - rules do not change during operation. There is no adaptation procedure to help the system improve over time. Time was not treated there as an important attribute of the system. Those systems were effectively used mainly for decision making on static data and static rules in static situations. Even though a module of a fuzzy neural network FuNN was developed as part of the FuzzyCOPE/2 (Kasabov, Kim et al, 97) which allows for rule extraction and adaptation, these features were not inherent in the developed systems.

In the next section a general framework of an ICM is introduced. It is designed to recognise patterns of events happening over different time scales. This is what a conscious mind does - it collects pieces of facts, information, data, signals over different time-scales (milliseconds, minutes, hours, years, etc) and recognises certain patterns (phonemes, words, sentences, heart-beat irregularities, moving objects from series of images, hand-written characters, long-term associations, etc.). Its applications in several areas are outlined. A connectionist realisation of this framework utilising fuzzy neural networks and spatial-temporal maps is presented in sections three, four and five. Sections three and four introduce also the main principles of fuzzy neural networks and spatial-temporal maps respectively. In section 5 some new cognitive and brain-like functions and principles of the human brain when processing multi-lingual spoken language information are discussed and a general framework of a multilingual spoken language recognition system is presented along with some preliminary experimental results. In several illustrative examples in this chapter data from the Otago Speech Corpus has been used (Kasabov et al, 95; Sinclair and Watson, 95). The Otago Speech Corpus on New Zealand English is available from the WWW: http://divcom.otago.ac.nz:800/COM/INFOSCI/KEL/speech.htm. Section 6 is a concluding section where current implementations of the framework are discussed along with directions for further research.

2. A Framework of an Intelligent Conscious Machine

Collecting information over different time-scales and forming meaningful patterns, structures, concepts, knowledge, to be further used and interpreted for inferring new knowledge, and for refining old one, is the aim of the framework presented here. The time-scales can be milliseconds, minutes, hours, years, etc. The patterns and structures can be phonemes, words, sentences, heart beat irregularities, moving objects, recognised from series of images, long-term associations, etc. Incoming information is grouped as it arrives over time and the best-matched existing structure is recalled or a new one is created. At a sub-conscious level the system recognises elementary events, sounds, and sequences of them. At the conscious level the system recognises meaningful patterns and structures, analyses its behaviour and improves in time. It adapts to new data, forms new structures, creates new associations.

A block diagram of the framework is given in fig.2.1. In general, each block consists of several modules working simultaneously, either on the same, or on different input data. Each module comprises many sub-modules, called elementary units.

Fig.2.1. A framework of an intelligent conscious machine for recognising complex patterns/ objects from time-series of events

The pre-processing units extract features and transform the raw input data according to certain time scales. The set of selected features is very important for the further operation of the whole system. Here a set of the two main formants is compared with a set of three time-lags of 26 element mel-scale vectors to represent speech data are compared and illustrated on fig.2.2 and fig.2.3.

In spite of the numerous publications on mapping phonemes into the feature space (map) of the first two main frequences (formants), we can easily proof that this space is far from being sufficient for unambiguous mapping. Figure 2.2 shows plots of four phoneme sounds taken from words spoken by one female and one male speakers of NZ English (speakers 12 and 17 from the Otago Speech

Corpus) in a two dimensional space of the first two formants. The plot illustrates the ambiguity at the elementary speech sound level for this particular feature space. By using the first two formants only, it is not possible to distinguish linguistically meaningful sounds, phonemes; e.g. it is hard to distinguish the female prononciation of /i/ and / /, the female /I/ and male / /, etc. At the same time the male and the female /e/, taken from 'get' appear at different places in the formant space as they differ in the second formant. It is necessary to use a higher level knowledge to be able to classify correctly close elementary sounds that have meaning in a spoken language, but even higher level knowledge may not be sufficient for the task.

Fig. 2.2. Plots of four phoneme sounds taken from words spoken by one female and one male speakers of New Zealand English (speakers 12 and 17 from the Otago Speech Corpus) in a two dimensional space of the first two formants. The plots contain ambiguous information.

Fig. 2.3. The male and the female realisation of /e/ from fig. 2.2 show similarity when shown as three 26 element-MSC vectors each of the vectors representing one 12 ms frame of the signal with 50% overlap between the frames (time-lags). The third graph shows the difference in intensity between the male and female pronunciation.

Ambiguity at the elementary sound level can be possibly reduced if another set of features is used, for example mel- scale coefficients (MSC) or mel-scale cepstrum coefficients, which filter the sound into predefined frequency band-filters. The male and the female /e/ from fig. 2.2 is shown in fig.2.3 as three 26-element MSC vectors each of the vectors representing one 12 ms frame of the signal with 50% of overlap between the frames (time-lags). The similarity between the male and the female patterns are visible in terms of frequency patterns. Of course there is difference in the intensity of the signals taken from the male and the female voice (here it is up to 20%).

In a general case, time is taken into account in the elementary event recognition module and several time-lags of the transformed input data are fed to the module where elementary events, patterns, are recognised with their corresponding matching degrees and confidence factors. The modules also learn and adapt to the most important features thus ignoring (possibly through forgetting) the unimportant ones at that time. They automatically find out and adapt to the optimal, for a particular elementary event, number and type of time-lags of the signal. In the realisation given in sections 3 and 5 fuzzy neural networks trained with forgetting are used to implement this block. Time is presented here as spatially ordered input features.

In the block of sequence of events recognition, the input vectors represent the recognised elementary events in the previous block, at a certain time interval, with their matching degrees and confidence factors. A sequence of such vectors activates possible referenced sequences (words). Some of the recognised sequences have meaning according to the concepts and structures defined in the following blocks, but some of them do not have meaning. Interaction between this block and the next ones is achieved trough the block of conscious decision making where meaning is identified. For example, a recognised sequence of sounds which represent a meaningful word in German or a meaningful word in Japanese, would not have a meaning to an English speaking person who does not speak German or Japanese. Defining the meaning of the speech sounds in terms of language is a conscious act. At each time (cycle) of recognition the interaction goes from the block of event-sequences recognition through the rest of the blocks to the highest level of the hierarchy and back to this block until a stable sequence or structure is recognised.

The feedback connection from the conscious decision making module "filters" all the perceived patterns, sequences of patterns and structures at all levels of the hierarchical framework. Only the meaningful ones proceed further on and are analysed and processed. Through learning and adaptation more and more perceived information patterns become meaningful and form short term or long term patterns depending on the time scale information from the time-scale recognition modules.

The conscious decision module makes adaptation possible at different levels, including the elementary events recognition level. New elementary events may be added to this module or some existing events may be split into sub-events, if this is needed for the recognition of meaningful concepts or structures.

Time is represented in the framework of fig.2.1 in many different ways, e.g. as spatially organised input vectors, as input-output associations in a module, as interaction between modules including feed-back interaction, or as symbols. Aggregating elementary events into larger and meaningful sequences, concepts, structures, is a challenging task. The general framework is illustrated in sections 3,4 and 5 with the use of fuzzy neural networks and spatial-temporal maps.

3. Fuzzy Neural Networks – A General Architecture and Applications for Phoneme Classification

3.1. The FuNN Architecture and its Functionality

Fuzzy neural networks are neural networks that realise a set of fuzzy rules and a fuzzy inference machine in a connectionist way (Yamakawa et al, 93; Hashiyama, Furuhashi, Uchikawa, 92; Jang, 93, Hauptmann and Heesche, 95; Kasabov, 96).

FuNN is a fuzzy neural network introduced first in (Kasabov, 96) and then developed as FuNN/2 in (Kasabov, Kim et al, 97). It is a connectionist feed-forward architecture with five layers of neurons and four layers of connections. The first layer of neurons receives the input information. The second layer calculates the fuzzy membership degrees to which the input values belong to predefined fuzzy membership functions, e.g. small, medium, large. The third layer of neurons represents associations between the input and the output variables, fuzzy rules. The forth layer calculates the degrees to which output membership functions are matched by the input data and the fifth layer does defuzzification and calculates values for the output variables. A FuNN has both the features of a neural network and a fuzzy inference machine. A simple FuNN structure is shown in Figure 3.1. The number of neurons in each of the layers can potentially change during operation through growing or shrinking. The number of connections is also modifiable through learning with forgetting, zeroing, pruning and other operations (Kasabov, 96; Kasabov and Kozma, 97; Kasabov, Kozma and Watts, 97).

The membership functions, used in FuNN to represent fuzzy values, are of triangular type, the centres of the triangles being attached as weights to the corresponding connections. The membership functions can be modified through learning as shown in Figure 3.2.

Several training algorithms have been developed for FuNN:

(a) A modified back-propagation (BP) algorithm that does not change the input and the output connections representing the membership functions.

(b) A modified BP algorithm that utilises structural learning with forgetting, i.e. a small forgetting ingredient, e.g. 10^{-5}, is used when the connection weights are updated (see Ishikawa, 96; Kozma et al, 96; Kasabov , Kozma and Watts, 97).

(c) A modified BP algorithm that updates both the inner connection layers and the membership layers. This is possible when the derivatives are calculated separately for the two parts of the triangular membership functions. These are

also the non-monotonic activation functions of the neurons in the condition element layer.
(d) A genetic algorithm for training (Kasabov, Kozma and Watts, 97).
(e) A combination of any of the methods above used in different time intervals as part of a single training procedure

Fig. 3.1. A FuNN structure for two initial fuzzy rules: R1: IF x1 is A1 (DI1,1) and x2 is B1 (DI2,1) THEN y is C1 (CF1); R2: IF x1 is A2 (DI1,2) and x2 is B2 (DI2,2) THEN y is C2 (CF2), where DIs are degrees of importance attached to the condition elements and CFs are confidence factors attached to the consequent parts of the rules (adopted from (Kasabov, 96)). The triplets (s,a,o) represent specific for the layer summation, activation, and output functions.

Fig. 3.2. Initial membership functions (solid lines) of a variable x (either input, or output) represented in a FuNN and the membership functions after adaptation (dotted lines). The boundaries, to which each centre can move but not cross, are also indicated.

Several algorithms for rule extraction from FuNN have been developed and applied (Kasabov, 96). One of them represents each rule node of a trained FuNN as an IF-THEN fuzzy rule as shown in fig.3.1.

FuNNs have several advantages when compared with the traditional connectionist systems or with the fuzzy systems:
(a) They are both statistical and knowledge engineering tools.
(b) They are robust to catastrophic forgetting, i.e. when further trained only on new data, they keep a reasonable memory of the old data.
(c) They interpolate and extrapolate well in regions where data is sparse.
(d) They can be used as replicators, where same input data is used as output data during training; in this case the rule nodes perform an optimal encoding of the input space.
(e) They accept both real input data and fuzzy input data represented as singletons (centres of gravity of the input membership functions))
(e) They are appropriate tools to build multi-modular IIS as explained next.

Fuzzy neural networks have been used so far for tasks of speech recognition. Some of the experiments use a large, single, neural network for classifying phonemes on their formant input values (Mitra and Pal, 95; Ray and Ghoshal, 97) and to extract fuzzy rules. Other experiments use hybrid neuro-fuzzy systems in a modular approach (Kasabov, 95,96; Kasabov, Kozma et al, 97a, 97b).

3.2. Using FuNNs for phoneme recognition

Here FuNNs are used to learn and classify phoneme data. Three 26-element mel-scale coefficients (MSC) vectors, representing the speech signal at three consecutive time frames of 12 ms each, are used as initial inputs.

Through training with forgetting, each FuNN unit is tailored to the specific phoneme (sound). After the training procedure and a consecutive pruning of the very small connections, only the important inputs that correspond to significant for the phoneme time-lags, and the important MSC, are kept in the FuNN structure. This is illustrated on fig. 3.3.

A FuNN structure is initialised as 78-234-10-2-1 and then trained with forgetting on both male and female data of the phoneme /e/, as positive data and the rest of the phoneme data as negative data. The training and testing data is taken from 139 words pronounced three times each by a male and a female speakers of NZ English, as explained in the Otago Speech Corpus, the male speaker being #12 and the female #17. The FuNN structure has been significantly simplified through training with forgetting and a consequent pruning. As it can be seen from the third figure on fig. 3.3 only three rule nodes have left. The condition element nodes and the left connections from them to the rule nodes, correspond to the main frequences of the phoneme /e/ realisation as shown on the top of fig. 3.3. The bright areas there show high energy of the signal for a particular MSC. It can also be seen that more connections from the first time-lag input vector are left which suggests a higher importance of this time-lag. The

trained /e/ FuNN, when tested on new data, showed correct true positive and true negative activation (the bottom figure on fig. 3.3)

Fig. 3.3. A FuNN used to classify the phoneme /e/ from one male and one female speech data: (a) A selected set of MSC vectors of the phoneme /e/ realisation (speakers #12 and #17 from the Otago Speech Corpus); each of the vectors represent three time lags (3x26=78 elements); (b) A FuNN trained to classify phoneme /e/ data without forgetting; (c) The same FuNN trained with forgetting; (d) Test accuracy of recognition – all phonemes in English extracted form spoken by the two speakers words are used for testing; the last part of the data are true phoneme /e/ realisation data.

3.3. FuNN-based Intelligent Multi-modular Systems.

Fig. 3.4a shows a block diagram of a FuNN-based IIS. It consists of a FuNN-based module which includes single FuNN-units for each class (elementary event, patter, etc.), a module for rule extraction and explanation, and a module for adaptation. Here, adaptation is the process of on-line training when a single

FuNN improves its performance based on observation and analysis how its performance compares to the reality. The modules for adaptation and explanation may be considered as forming the conscious decision making module as shown in fig. 2.1. Adaptation in a multi-modular FuNN structure is based on individual tuning of single FuNN-units if the analysis of the performance of the whole system shows that those are reasons for unsatisfactory performance or points of improvement. One scheme for adaptation is shown in fig. 3.4b where a copy of a FuNN is trained on-line to improve its performance while the main FuNN-unit is in operation. After a certain time interval the copy substitutes the main FuNN unit.

Fig. 3.4. (**a)** A multi-modular FuNN-based Intelligent Information System. (**b**) A scheme for adaptation in a multi-modular FuNN-based IIS

3.4. Multi-modular FuNN-based Systems for All-Elementary Event (All-Phoneme) Classification

Building FuNN-based IIS is illustrated here on the whole set of phoneme data of one male and one female speaker (#12 and #17) comprising 10,000 training examples and 5,000 validation examples. A single FuNN is trained to classify speech input data into one of the 43 phonemes in New Zealand English (phonemes #44 and #45 are not included), in the same way as it was explained above on the case of phoneme /e/. Figure 3.5 shows validation results. The results are satisfactory for most of the phonemes. Obviously further tuning of the phonemes ## 24, 30, 39, 43 is needed which is possible because of the modular approach taken. All the phoneme FuNNs will be indeed trained further on other speakers' data. The modular approach also allows for adaptation of individual phoneme FuNNs to new speakers, accents, dialects. New phoneme FuNNs can be easily added if necessary along with the modification of the existing ones. The trained FuNN-based all-phoneme classifier is shown as part of a multilingual spoken recognition system in section 5. The overlapping between the phoneme classification can be shown in a form of a confusion matrix as shown in fig. 3.6. The matrix represents the winner takes all principle when the number of activations (correct and wrong) of each phoneme FuNN is counted and

presented as a level of darkness. This matrix suggests way to measure similarity in sounding between different phonemes.

Fig. 3.5. Preliminary evaluation of the test accuracy of the phoneme recognition in the FuNN units across all the phonemes in New Zealand English. FuNNs have been trained and tested on one male and one female speakers data from the Otago Speech Corpus. The black and white bars represent the true positive and the true negative classification accuracy resepectively.

Fig. 3.6. A confusion matrix of the validation classification of NZ English phoneme data in the FuNN-based multi-modular all-phoneme classifier on one male and one female data

4. Spatial-Temporal Maps: A general Introduction and Applications for Mapping Elementary Events (Phonemes) and Sequences of Events (Words)

4.1. A General Introduction

Spatial-temporal maps (STM) are connectionist structures which have time sequence of vectors as inputs and a topologically, spatially organised map as an output. Kohonen self-organised maps (SOM) (Kohonen, 90) with time sequence of vectors as inputs, are examples of STM. Figure 4.1. represents a block diagram of a STM. STM can be trained either in a self-organised, unsupervised way, or in a supervised one.

Fig. 4.1. A block diagram of a STM.

STM can be used to map similar sequences of elementary events over time intervals (not necessarily equal in duration) into topologically close areas on the map. When used in the general framework from fig. 2.1, to realise the block of sequence of events recognition, the STM may recall, after new data is input, both meaningful and meaningless sequences, which are further defined by the conscious decision making block through a feed-back connection.

4.2. STM for Mapping Phoneme- and Word- Data

One of the first applications of the Kohonen SOM was for phoneme and word recognition. In this section Kohonen SOMs are used to map temporal sequences of so called phoneme activation vectors into a dictionary of words. Mapping phonetic representation of words into a *"sounds-like" STM* is a new approach introduced and used here. It allows for storing, updating and retrieving words from large dictionary. The output of the module of elementary sounds (phoneme) recognition is an n-element activation vector produced every time frame (say 6ms). The activation vector contains the activation of the elementary events (phonemes) at a certain time frame. In the case of phoneme recognition, this is a phoneme activation vector (PAV). A sequence of PAVs is further aggregated in a shorter sequence of PAVs which is then mapped in a dictionary of words represented as a trained STM. Mapping PAV is illustrated below for the NZ English phonemes and for a small set of English words.

Synthetic PAVs can be created based on the expected similarity between the sounding of the phonemes (see for example the confusion table from fig.3.6 and the acoustic features given in the appendix between the phonemes, e.g. alveolar, etc.). A PAV contains a value of activation of 1 for that phoneme, lesser activation values for similar phonemes and values of 0 for different phonemes (Kasabov, Kozma, Kilgour et al, 97a; 97b). Mapping the PAVs into a SOM is shown in fig. 4.2. The SOM was trained with 43 phoneme activation vectors for 10,000 epochs. The phonemes are clearly distinguished on the map. Further training and adaptation of the SOM on more data is possible.

Fig. 4.2. Mapping the synthetic phoneme activation vectors of the phonemes in NZ English into a SOM. The SOM was trained with 43 phoneme activation vectors

Through using PAVs the phonetic transcription of the words can be used to map all the words from a dictionary (regardless of its size, e.g.. 2,000 or 200,000) into a STM of "sounds-like" words. For example, a phonetic transcription of the word 'pat' can be represented as 3 times 43-element PAVs. Figure 4.3 shows the SOM for several English words.

Fig. 4.3. The "sounds-like" word SOM for several English words. Inputs are synthetic PAVs. It is seen on the map that similar PAVs activate neurons in the same area. The activation of all the output neurons is shown and the winning neuron is marked as "*".

Representing language dictionaries as STM and SOM in particular, have several advantages when compared to the traditional database representation:
(a) it accounts for similarities between the words in the dictionary;
(b) it makes the whole process of searching through a dictionary effective regardless of the size of the dictionary as it is achieved by one recall procedure through the STM;

STM can be used for representing higher level modules form the framework of fig.2.1, e.g. the concept recognition modules etc. as shown in section 5. Both the FuNN-based phoneme classifier and the STMs are used for a partial implementation of a multi-lingual spoken recognition system in the next section.

5. A Framework for Multi-lingual Speech Recognition Systems

5.1. The Problem of Speech Recognition and the Role of Consciousness. A Framework of an ICM for Spoken Language Recognition

Spoken language recognition and understanding in computer systems is a challenging task (Cole et al, 95; Altman,90; Jusczyk,97). The aim of it is two-fold:
(a) as the best machine for this task is the human brain, the task will stimulate and will require further study on the speech perception and language learning in humans;
(b) as the task involves intelligence and consciousness, even partial solutions of the task will bring useful brain-like computing methods, new methods of cognitive engineering and therefore new frameworks for building IIS and ICM.

The task has two main phases, namely sub-conscious, i.e. the phase of sounds and the sequences of them (words) recognition regardless of their meaning, and conscious - the phase of speech sounds, words, sentences etc. recognition in terms of language (or languages). The task involves time at several scales, e.g. milliseconds in terms of elementary sounds (phonemes), seconds in terms of words, minutes or longer periods in terms of sentences and logical associations between their meanings. This task fits well the general framework explained in section 2 and presented in fig. 2.1, as it is shown in fig. 5.1.

The linguistic, conscious decision making block in the framework from fig. 2.1 has a significant role in the whole process of spoken language recognition. Applying consciousness and language awareness is the only away to deal with the tremendous variability and ambiguity in speech. This is the way for a system to deal with problems such as:
(a) Adapt to new accents and dialects through applying linguistic knowledge about their relationship with some already learned ones;
(b) Distinguish close sounds through the context of a language at the higher level of information processing which information is fed back to the low level processing through the feedback from the conscious decision making block;
(c) Acquire new language, thus turning some of the recognised meaningless sounds and words into meaningful ones.

Fig. 5.1. A framework of an ICM for spoken language recognition

A partial realisation of the framework from fig.5.1 for the case of multi-lingual system is given in fig. 5.2. It uses FuNNs for the phoneme recognition and STMs for word and concept recognition.

5.2. Multi-lingual Spoken Language Recognition

Learning a second and other languages has been investigated in several papers and books (Zatorre, 89; Juszcyk, 97). Recently, fMRI (functional Magnetic Resonance Imaging) of the activation of neurons at particular spatial areas of the Broca's area of the cortex, when words of different languages were spoken to a person who speaks these languages, was experimented (Kim et al, 97). The investigation shows that when two or more languages are learned at an early age (under the age of 7) the activation centres which respond highly to pronounced words in two or more learned languages, are in the same Broca's area (a distance of about 1mm has been measured). When a second language is learned at a later age the activation centres of the two languages belong to different areas of the Broca's area (a difference of about 8 mm has been measured). Broca's area is known as a phonetically sensitive, anterior language area. At the same time another language sensitive area of the brain, Wernicke's area, which is a language posterior area, allocates same centres for all the multiple languages regardless of the age of acquisition. It is concluded there that learning several languages at an early age makes use of the same region in the cerebral cortex for the elementary sounds and the sounding of the words of these languages.

The above referenced research and some other investigations of the spatial-temporal mapping of the speech and language in the human brain support the approach taken here, and graphically illustrated in fig. 5.2, for building multi-lingual ICMs for spoken language recognition. The elementary sound recognition module, the "sounds-like" word module and the concept recognition module are shared between the languages recognisable in the system. The language structure

recognition modules are different for the different languages and they give indication to the conscious decision making module whether the pronounced sounds and words belong to the languages the system can understand, so it is a language-aware system. The FuNN-based phoneme classifiers and the STMs described in the previous sections are used for a partial implementation of the multi-lingual spoken recognition system as shown in fig.5.2.

Fig. 5.2. A schematic diagram of a multi-lingual spoken recognition system illustrated with the process of recognition of the word ‘cat’ in six languages.

The low level FuNN-based elementary sounds module and the STM for sequences of sounds recognition module are shared between all the languages used in the system. They are expandable to include new elementary sound recognition units if such are needed for a new language. The STM module for recognising the concepts and the meanings of the words and sentences is also shared as humans share common sense regardless of the language they speak.

Words in different languages that have same meaning would be mapped at the same point (area) in the map of concepts, e.g. the concept of “pets”. Of course learning new languages would lead to expanding the concept STM with new concepts. The higher level of language structure recognition and the awareness about the languages ‘known’ to the system is a conscious process done in a close interaction between the last block from fig. 5.1 and fig. 5.2 and the conscious decision making block.

After the elementary sounds are recognised, n-element PAVs, containing the aggregated activation of each of the n phonemes over a certain period of time, are fed to the sounds-like word map. This STM has k-inputs, each of them being n-element phoneme activation vector. The sounds-like map receives feedback from the higher level modules in the system in order to refine the choice of a group of words to be further processed at a higher level. These maps are trained on both artificial data generated from linguistic knowledge (see the STM in section 4) and real data - the data from the previous module on real data inputs. Figure 5.3 shows a sounds-like SOM of words from six languages (English, Maori, German, Russian, Bulgarian and Japanese). The 7x7 SOM was trained with the PAVs of these words each of them represented as seven 43-element PAVs.

Fig. 5.3. A sounds-like SOM which maps words from six languages. The SOM was trained with the phoneme activation vectors of these words.

6. Conclusions and Directions for Further Research

This chapter discusses some issues in the area of intelligent information systems and suggests a general framework of an intelligent conscious machine along with its application to multi-lingual spoken recognition systems. The two levels of operation of the framework are lower, sub-conscious, where sounds and their

sequences are recognised, and higher, conscious, where meaningful words, concepts and language structures are recognised. The block of conscious decision making feeds information back to all the modules of the framework thus realising adaptation and deliberate learning of new knowledge in the system, which can be extended to include new languages. A partial implementation and preliminary results are shown when fuzzy neural networks and spatial-temporal maps are used to realise some of the modules of the framework. A language dictionary is represented as a connectionist self-organised map which makes the search through the dictionary quick. The multi-lingual approach suggested in the chapter is in coherence with new evidence about speech and language mapping in the human brain of multi-lingual subjects. This research will continue towards using the framework for continuous speech, multi-lingual recognition systems.

Future research has been also planned towards applying the framework to other tasks, such as moving objects recognition, heart-beat variability estimation, fruit growth prediction.

Acknowledgements

This work was done as part of the UOO606 project funded by the PGSF of the FRST of New Zealand. The following colleagues contributed to some of the experiments presented here: Dr Robert Kozma, Richard Kilgour, Mark Laws.

References

1. Alexander, I. (1997) *Impossible Minds*, Imperial College Press
2. Altman, G. (1990) *Cognitive Models of Speech Processing*, MIT Press
3. Arbib, M. (1987) *Brains, Machines and Mathematics*, Berlin. Springer Verlag.
4. Arbib, M. (ed) (1995) *The Handbook of Brain Theory and Neural Networks.* The MIT Press
5. Cole, R. et al (1995) The Challenge of Spoken Language Systems: Research Directions for the Nineties, *IEEE Transactions on Speech and Audio Processing*, vol.3, No.1, January 1995, 1-21
6. Handelman, D.A., Lane, H.S. and J.J. Gefland (1990) Integrating Neural Networks and Knowledge-Based Systems for Intelligent Robotic Control, *IEEE Control Systems Magazine*, Vol.10, No.3,
7. Hashiyama, T., Furuhashi, T., Uchikawa, Y.(1992) A Decision Making Model Using a Fuzzy Neural Network, in: *Proceedings of the 2nd International Conference on Fuzzy Logic & Neural Networks*, Iizuka, Japan, 1057-1060.
8. Hauptmann, W., Heesche, K. (1995) A Neural Net Topology for Bidirectional Fuzzy-Neuro Transformation, in*: Proceedings of the FUZZ-IEEE/IFES*, Yokohama, Japan, 1511-1518.

9. Hendler, J. and L.Dickens (1991) Integrating Neural Network and Expert Reasoning: An Example, in: *Proceedings of AISB Conference*, eds.Luc Steels and B.Smith, Springer Verlag, pp.109-116.
10. Ishikawa, M. (1996) Structural Learning with Forgetting, *Neural Networks*, 9, 501-521.
11. Jang, R. (1993) ANFIS: adaptive network-based fuzzy inference system, *IEEE Trans. on Syst.,Man, Cybernetics*, 23(3), May-June 1993, 665-685
12. Jusczyk, P. (1997) *The Discovery of Spoken Language*, MIT Press
13. Kasabov, N.(1996) *Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering*, The MIT Press, CA, MA.
14. Kasabov N. (1995) Hybrid Connectionist Fuzzy Rule-based Systems for Speech Recognition*", Lecture Notes in Computer Science/Artificial Intelligence*, Springer Verlag, No.1011, 20-33
15. Kasabov, N. (1990) Hybrid Connectionist Rule Based Systems. in: Ph.Jorrand and V.Sgurev (Eds). *Artificial Intelligence -Methodology, Systems, Applications.* North-Holland, pp.227- 235
16. Kasabov, N. (1993) Hybrid Connectionist Production Systems, *Journal of Systems Engineering*, vol.3, No.1, pp.15- 21, Springer Verlag London.
17. Kasabov, N. and Kozma, R. (1997) Adaptive fuzzy neural networks and applications for chaotic time-series analysis and phoneme-based speech recognition, *IEEE Transactions on Neural Networks*, to appear
18. Kasabov, N., Kilgour, R., Sinclair, S. (1997) From hybrid adjustable neuro-fuzzy systems towards connectionist-based adaptive systems for phoneme-, word-, and language recognition. *Fuzzy Sets and Systems*, to appear
19. Kasabov, N. (1995) Hybrid Connectionist Fuzzy Production Systems - Towards Building Comprehensive AI, *Intelligent Automation and Soft Computing*, 1:4, 351-360
20. Kasabov, N., Kim J S, Watts, M., Gray, A (1997) FuNN/2- A Fuzzy Neural Network Architecture for Adaptive Learning and Knowledge Acquisition, *Information Sciences - Applications*, 1997, in print
21. Kasabov, N., Kozma, R. and Watts, M. (1997) Optimisation and Adaptation of Fuzzy Neural Networks Through Genetic Algorithms and Structural Learning , *Information Sciences – Applications*, in print
22. Kasabov, N., Sinclair, S., Kilgour, R., Watson, C., Laws, M. and Kassabova, D. (1995) Intelligent Human Computer Interfaces and the Case Study of Building English-to-Maori Talking Dictionary, in: *Proceedings of ANNES'95*, Dunedin, IEEE Computer Society Press, Los Alamitos, 294-297
23. Kasabov, N., Kozma, R., Kilgour, R., Laws, M., Taylor, J., Watts, M., Gray, A. (1997a) HySpeech/2: A Hybrid Speech Recognition System, *Proceedings of the ICONIP/ANZIIS/ANNES'97*, Dunedin, 24-28 November 1997, Springer Verlag, Singapore
24. Kasabov, N., Kozma, R., Kilgour, R., Laws, M., Taylor, J., Watts, M., Gray, A. (1997b) Speech Data Analysis and Recognition Using Fuzzy Neural Networks and Self-Organised Maps, in: Kasabov, N. and Kozma, R. (Eds) *Neuro-Fuzzy Tools and Techniques*, Physica Verlag, Heidelberg, in print

25. Kim, K., Relkin, N., Lee, K.-M., Hirsch, J. (1997) Distinct Cortical Areas Associated with Native and Second Languages, *Nature*, vol.388, July, 171-174
26. Kohonen, T. (1990) The Self-Organizing Map. Proceedings of the IEEE, vol.78, N-9, pp.1464-1497.
27. Kozma, R., Sakuma, M., Yokoyama, Y., Kitamura, M. (1996) On the accuracy of mapping by neural networks trained with backpropagation with forgetting, *Neurocomputing*, 13, 295-311
28. Mitra, S., Pal, S. (1995) Fuzzy Multi-Layer Perceptron, Inferencing and Rule Generation, *IEEE Transactions on Neural Networks*, vol.6, No.1, 51-63
29. Ray, K., Ghoshal, J. (1997) Neuro-Fuzzy Approach to Pattern Recognition, *Neural Networks*, vol.10, No.1, 161-182
30. Sinclair, S., Watson, C. (1995) Otago Speech Data Base, in: Proceedings of ANNES'95, Dunedin, IEEE Computer Society Press, Los Alamitos
31. Takagi, H. (1990) Fusion Technology of Fuzzy Theory and Neural Networks - Survey and Future Directions, in: Proc. First Int. Conf. on Fuzzy Logic and Neural Networks, Iizuka, Japan, July 20-24, pp.13-26.
32. Yamakawa, T., Kusanagi, H., Uchino, E. and Miki, T.(1993) A new Effective Algorithm for Neo Fuzzy Neuron Model, in: *Proceedings of Fifth IFSA World Congress*, (1993) 1017-1020.
33. Zadeh L. (1984) Making Computers Think Like People , *IEEE Spectrum*, Aug 1984, pp.26-32
34. Zadeh, L. 1965. Fuzzy Sets, *Information and Control*, vol.8, 338-353.
35. Zatorre, R. (1989) On the representation of multiple languages in the brain: Old problems and new directions, *Brain and Languages*, 36, 127-147

Appendix. The phonemes in NZ English from the Otago Speech Corpus

No	ASCII	Example	Class	Characteristics		
01	p	pin	Consonant	Bilabial	Plosive	Unvoiced
02	b	bay	Consonant	Bilabial	Plosive	Voiced
03	t	toy	Consonant	Alveolar	Plosive	Unvoiced
04	d	die	Consonant	Alveolar	Plosive	Voiced
05	k	key	Consonant	Velar	Plosive	Unvoiced
06	g	get	Consonant	Velar	Plosive	Voiced
07	f	five	Consonant	Labiodental	Fricative	Unvoiced
08	v	van	Consonant	Labiodental	Fricative	Voiced
09	T	thick	Consonant	Dental	Fricative	Unvoiced
10	D	then	Consonant	Dental	Fricative	Voiced
11	s	see	Consonant	Alveolar	Fricative	Unvoiced
12	z	zink	Consonant	Alveolar	Fricative	Voiced
13	S	ship	Consonant	Palato-alveolar	Fricative	Unvoiced
14	Z	measure	Consonant	Palato-alveolar	Fricative	Voiced
15	h	he	Consonant	Glottal	Fricative	Unvoiced

16	tS	chin	Consonant	Palato-alveolar	Affricate	Unvoiced
17	dZ	jam	Consonant	Palato-alveolar	Affricate	Voiced
18	m	me	Consonant	Bilabial	Nasal	Voiced
19	n	not	Consonant	Alveolar	Nasal	Voiced
20	N	sing	Consonant	Velar	Nasal	Voiced
21	l	light	Consonant	Alveolar	Approxim.	Voiced
22	r	ring	Consonant	Post-alveolar	Approxim.	Voiced
23	w	win	Consonant	Velar	Approxim.	Voiced
24	j	yes	Consonant	Palatal	Approxim.	Voiced
25	I	sit	Monophth.	Close	Front	Unround.
26	e	get	Monophth.	Mid	Front	Unround.
27	&	cat	Monophth.	Open	Front	Unround.
28	V	hut	Monophth.	Open	Central	Unround.
29	A	hot	Monophth.	Open	Back	Rounded
30	U	put	Monophth.	Mid	Back	Rounded
31	i	see	Monophth.	Close	Front	Unround.
32	a	father	Monophth.	Open	Back	Unround.
33	O	sort	Monophth.	Close	Back	Rounded
34	3	bird	Monophth.	Mid	Central	Rounded
35	u	too	Monophth.	Close	Back	Rounded
36	eI	day	Diphthong	Closing		
37	ai	fly	Diphthong	Closing		
38	Oi	boy	Diphthong	Closing		
39	OU	go	Diphthong	Closing		
40	aU	cow	Diphthong	Closing		
41	i@	ear	Diphthong	Centring		
42	U@	tour	Diphthong	Centring		
43	e@	air	Diphthong	Centring		
44	Q	silence				
45	@	banana	Monophth.	Mid	Central	Unround.