Developing a Model of Speech Recognition Process from an Autopoietic Approach

Liew Eng Siang and Abd Manan Ahmad
Fakulti Sains Komputer & Sistem Maklumat
Universiti Teknologi Malaysia
81310 UTM Skudai, Johor
Email: esliew@hotmail.com

Abstract– **Speech recognition is a time-tested research into the possibilities of reproducing human like qualities in machines such as computers. In the 40 years of exploration in this field, there are many milestones that have been set. Among the recent advances in speech recognition, the deployment of statistical techniques like Hidden Markov Models and connectionist techniques like Time Delayed Neural Network have rendered speech recognition into a very mature and commercially viable venture. Therefore, the direction of research in this paper is aimed at the improvement of current techniques rather than reinventing the wheel; particularly, autopoiesis is introduced in this paper to give a flair of self-organization to the cognitive process of speech recognition. It is hoped that this paper would provide complimentary alternatives to speech recognition rather than exclusive solutions it.**

I. INTRODUCTION

The art of human communication has been honed to almost perfection through the millennia of man's progress in civilization. However, this cannot be said about the human-computer interaction. The basic interface of man and machine is still very much in a primitive mode. The keyboard, mouse and scanner are still the most important input devices. What about the microphone and the digital camera? Can all these devices ever be made into input devices that play a more important role apart from being just data streaming devices? What of cognition of the computer with the aid of these devices? These and many other issues are the targets of this paper.

Speech recognition has gained a strong foothold in the world of research. However the autopoietic approach has yet been explored fully as self-learning and self-organizing algorithms are still very much in the stages of experimental development. The term self-organization has become a battle standard for championing phenomena that appear to determine their own forms and processes. There is now a widespread interest in applying theories of self-organization to analysis and reengineering of non-linear systems. Therefore, it is interesting to note that autopoiesis bears a very promising future for further research and development work.

II BACKGROUND ON SPEECH

Human speech is basically an acoustic signal produced by the human vocal mechanism. The workings of this mechanism depend solely on the differential in air pressure of the chest cavity. Voiced sounds of speech are actually produced by the vibratory action of the vocal cords [1]. This vibration is the effect of the oscillation of the cord due to the air flowing through the vocal tract and the local pressures that this air flow causes. Therefore, human speech could be easily analyzed from the aspects of frequency, amplitude, harmonic structure and resonance of the oscillation produced [2].

Frequency and amplitude refers to the physical nature of the acoustic signal of speech. Frequency is the rate at which the signal oscillates and is measured in Hertz (Hz); or cycles per second. Amplitude, on the other hand, refers to the strength, or loudness, of the signal propagated across a medium and is measured in decibels (dB). These two measurements of speech would form the basis of the speech recognition model proposed.

2.1 Basics of Speech Recognition

Formulation of a generic speech recognition system would require the primary components of an input mechanism, a signal processing engine, a cognitive mechanism, and an output device. These basic components reflect the nature of the recognition process as one of a reductionist paradigm; in that, the sum of the parts make the whole of the system. However, if a more holistic approach were to be introduced, the whole development process of such said system would hardly be an empirical, methodological, and scientific study. Therefore, the model proposed in this paper would be a constitution of components that make a speech recognition system and the emphasis of research from this paper would be on those components that benefit most from the attention of the autopoietic theory.

The input mechanism receives the raw speech for the system and prepares this raw input for further processing from the signal-processing engine. In general, most speech recognition system samples the raw speech signal at a rate between 4 kHz and 20 kHz [3] [4] [5]. The speech is then transformed to frames at intervals of 10 or 20 ms for the purpose of simplification of contributory factors and

compression of the signal produced. This transformation is done by the signal-processing engine and it involves techniques such as Fast Fourier Transform (FFT), Perceptual Linear Prediction (PLP), and Linear Predictive Coding (LPC) [4] [5].

From the speech frames generated, cognition of the frames would involve two aspects; i.e. the acoustic variability and the temporal variability [4]. The acoustic variability refers to the difference in pitch, volume, pronunciation, and so forth. The current approaches to handling cognition from the acoustic variability point of view are the template-based approach, the knowledge-based approach and the statistical-based approach. On the other hand, temporal variability refers to the rate of speech and has been proven that dynamic time warping algorithm is the solution to the issue. The output of this recognition system can be directly or indirectly applied to the areas of robotic control, machine translation, automated dictation and other language dependent systems.

III AUTOPOIESIS AND COGNITION

Autopoiesis is the study into the essence of the living. It was developed as an explanation into the workings of life from an organizational perspective of cognition. The theory was introduced by two Chilean biologist; i.e. Humberto R. Maturana and Francisco J. Varela. It was meant to designate the organization of a minimal living system [6]. This theory attempts to put answers, beyond the point of philosophical discussions, to the questions of what it is to be alive and what differentiates a living entity from that of an automated machinery.

However, the focus of this paper would be upon the cognitive aspect of the theory and the self-organizational prospects that the theory brings to the development of a speech recognition system. After all, 'cognition is a biological phenomenon and can only be understood as such; any epistemological insight into the domain of knowledge requires this understanding' [7]. Therefore, it is in cognition that the answers to a speech recognition system will be found. It is also the aim of this paper to bring to attention the significance of autopoiesis in the realm of learning and cognitive processing.

3.1 An Introduction to Autopoiesis

Autopoietic systems in general are defined in terms of their organization [8]. The main concerns of autopoiesis is on organization and structure because 'the organization of a machine (or system) does not specify the properties of the components which realize the machine as a concrete system, it only specifies the relations which these must generate to constitute the machine or system as a unity' [7]. Therefore, the organization of a system is independent of the properties of its components though a given system can be realized by many different structures. However, for a system to have a concrete constitution in any given space, actual components must be defined in that space and have the properties which allow them to generate the relations that had define it, hence the importance of structure.

3.2 Complimentary of Structure and Organization

Notions of structure and organization are akin to ideas of the fundamentals of matter. The structure of matter consist of atomic particles, that are, themselves, a constitution of sub-atomic particles. Whereas, organization refers to the invisible laws of nature that binds all these structure together to form the matter. However in biology, organization refers to the group identity of living entities; i.e. the pattern of relationships among their components which bear resemblance in many of their significant characteristics. On the other hand, structure refers to the particulars within a given entity, i.e. the physical properties of the components and the roles of the components play in the making up of the whole entity.

Therefore, the complimentarity between organization and structure is reflected upon the cooperation of both concepts in maintaining a balanced and smooth relationship known as life. This is due to the fact that organization can only exist in terms of relationships amongst structures and structure exists only in filling the roles in those relationships. The simplicity of this relationship of organization and structure highlights the underlying complexity of what would be termed alive for there must exist an inexplicable mechanism that regulates the workings of this relationship. It is this mechanism that would provide the key to achieving a self-organizing system that we could term alive.

3.3 Operational Closure

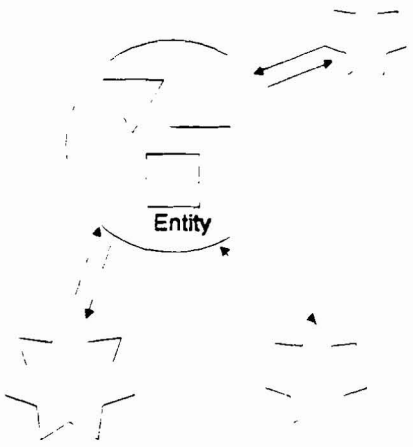

Figure 1 shows an entity that interacts with the environment yet the external environment doesn't affect the internal structure

Closure is a systems notion that refers to the containment of a system's operation within a system's boundary. For example, living entities are

open to the environment in terms of energy and material but are operationally close because the physical boundaries with which a living entity is confined to has no correlation between the internal structure of the entity and that of the external view of its environment. However, this phenomenon is a physical representation of the relationship of a living entity with its environment. It doesn't portray the whole picture of what it means to be alive.

3.4 Structural Coupling

Structural determination describes the actual course if change in a systemic entity is guided by the entity's own structure rather than influences of the entity's environment. Thus, given this principle, interaction among systems can be summed up as 'a history of recurrent interactions leading to the structural congruence between two systems' [9]. Structural coupling is the label for ongoing engagements between systems that positively return a change in structure of each entity. Therefore, structural coupling describes ongoing mutual co-adaptation without hinting of a transfer of some ephemeral force or information across the boundaries that separate the engaged systems.

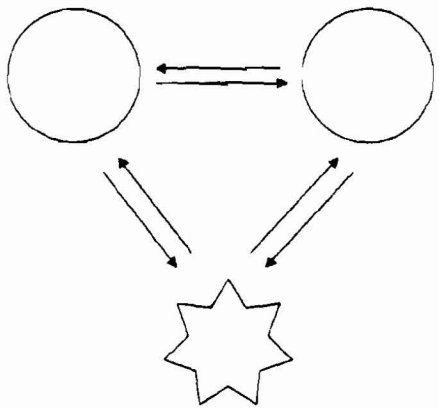

Figure 2 shows that two entities interacting with one another while maintaining their interaction with the environment

3.5 Observation and Cognition

From the autopoietic point of view, we, as living entities, are too the observers of our lives. We make these observations in a communicative process and try to explain these observations in a conversational methodology. The reality derived from our observations could be divided into two distinct forms; i.e. an objective reality and a personal reality. Objective reality provides us with a referral point from which we see ourselves from an external point of view; much like as a third person view of ourselves. Personal reality however, refers to the assessments made of the derived objective reality perceived by the observer; i.e. ourselves.

Therefore, achieving cognitive status for a living entity is but accumulating numerous samples of observations that are commonly known as experience. To be endowed with the cognition title, the living entity must benefit from cognition by being capable of change to its awareness in the wake of an ever evolving environment and this can only arrive from a properly constructed internal structure that could facilitate the storage of all foresaid experiences. Thus, cognition is contingent on embodiment because it is truly a consequence of the entity's specific structure.

IV SELF-ORGANIZING MAPS: A CASE STUDY

Kohonen had developed a phonetic typewriter for the Finnish language [10]. The typewriter receives the input in the form of speech and converts the speech into text. Kohonen's typewriter is a speaker-dependent speech recognition system for continuous speech with unlimited vocabulary [11]. However, isolated word recognition could be produced too, through this typewriter. The dimensions achieved by the typewriter was largely facilitated by the criteria of the language used, in this case Finnish, for the Finnish language, that is highly phonetic, enables the decomposition of the speech recognition problem into two key issues; i.e. the recognition of individual phonemes and the translation of phonemes to letters.

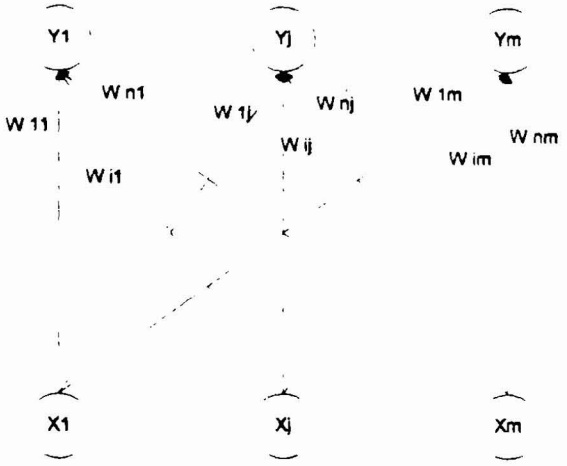

Figure 3 depicts the basic organization of the Self-Organizing Map with its inputs $X_1, ..., X_j, ..., X_m$ and outputs $Y_1, ..., Y_j, ..., Y_n$

4.1 The System Description of KSOM

Kohonen's Self-Organizing Map (KSOM) is an Artificial Neural Network (ANN) that is unique in that it is based on the unsupervised topographic mapping network and has been commercialized to some extent [11]. The underlying structure of the system is based upon the phonemes of the Finnish language. Phonemes form a topographic map of the network's weights through self-organization. Detection and activation of a phoneme is achieved through the mapping of the input to the closest prototype vector of the phoneme map that has been trained previously. The whole process of matching and recognition consist of a sequence of procedures.

First, a speech signal is sampled every 9.83 ms at 256 evenly distributed points. Then the sampled

signal is transformed using the Fast Fourier Transform (FFT) into frames with spectral power values that are further grouped into 15 spectral channels from 200 Hz to 5 kHz. The outputs from the 15 channels form 15 coefficients of the pattern vector. Kohonen used a two-dimensional array of 96 units with center-on surround-off lateral connections as in the Kohonen Map discussed in Section 4.2. Each unit then receives 15 connections from the 15 channels and the map begins to self-organize.

After stabilizing the self-organization, it is found that different examples of the same phoneme tend to cluster together on the map. Therefore, it would appear that units that cluster together seem to be of the same phoneme. This evolutionary process is the result of a continuous presentation of pattern vectors into the network thus forming a more static mapping of the phonemes. The process could be achieved because each pattern vector represents only a segment of signal shorter than the true phoneme. Therefore, any given test input would generate a path of phonemes through the map.

4.2 Formalism of KSOM

KSOM is a topology-preserving map in that it preserves the structure of the network while updating the interconnections between neurons and also while the process of learning is conducted. The learning in KSOM is a competitive, winner-takes-all strategy that only connections to winning neurons are updated and updating of the weights does not rely in any way on the spatial relations among the units in the competition layer of the network [11][12]. The two key mechanisms of this self-organizing map is that there exist a winning unit (neuron) that best responds to a given input and that there is modification to the connections of the winning unit to its neighboring units in the map.

The neighborhood within which units are updated together with the winning unit is known as the winning neighborhood. Basically, the winning unit and its neighborhood of interconnected units are more likely to respond to inputs similar to current inputs. To ensure that the neighborhood of units yields nonzero outputs, all the units should have their connections updated so that they are more likely to respond well to the same input. A simple way to achieve this is by modifying the lateral connections of the output layer.

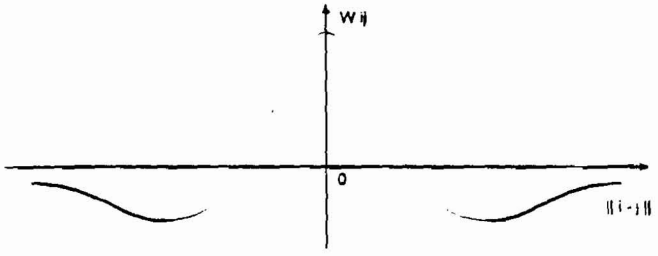

Figure 4 is a graph that shows the continuous values of the weights W_{ij} that correspond to a neighborhood function of $\|i-j\|$

V SPEECH RECOGNITION MODEL

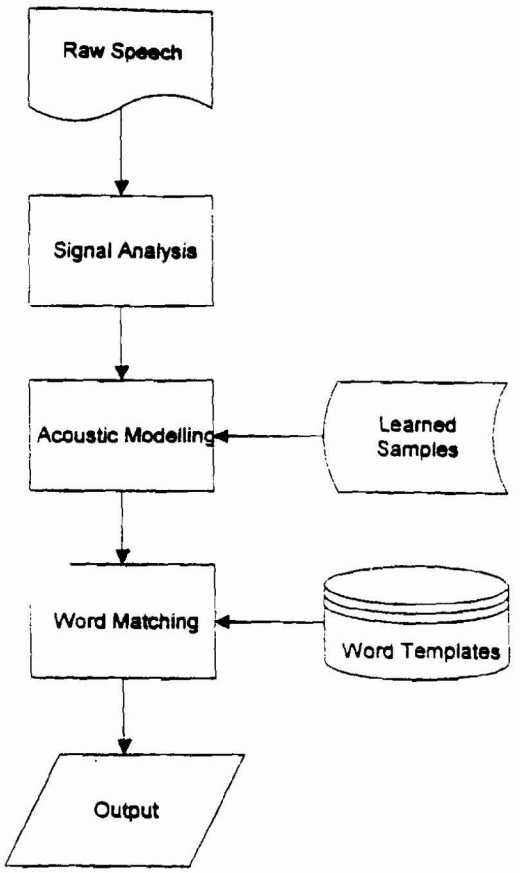

Figure 5 shows the preliminary design of the speech recognition system.

Basically, the model of the speech recognition has three main components; namely a signal analysis engine, an acoustic modelling mechanism and a word matching function. The signal analysis component would serve as the conversion function of the raw speech signal into a usable speech frame with the help of the Fast Fourier Transform (FFT) technique.

The second component, the acoustic modeller, would benefit most from the implementation of the autopoietic theory. To understand this implementation, let's assume that ε is the set of all the phonemes in a particular language. Then assume that A is a word

$$A = \{a_i : a_i \in \varepsilon, i \in \aleph^-\} \quad (1)$$

Suppose that S is the global set of speech that will be recognized, then

$$S = \{ A_i : i \in \aleph^+\} \quad (2)$$

If an input speech $X = [\varepsilon]^k, k \in \aleph^+$

$$X = \{ x_i : x \in [\varepsilon]^k, k \in \aleph^+\} \quad (3)$$

Therefore, a learning function $\Psi(x)$ would plot

$$\Psi(x_i) = \{ y_i, \forall y \in S\}, i \in \aleph^+ \quad (4)$$

The validation of this learning would come in the form of

$$\Psi'(y_i) = \{ a_i, \ni a_i \in S\}, i \in \aleph^+ \quad (5)$$

The formulation of Equation (5) could only be achieved with the help of introducing a function based on the autopoietic theory.

The word matching function of this model would be based on the Dynamic Time Warping (DTW) algorithm. The speech frame that has been identified would then be made to match with the words in the trained database or else, if it is a new word input, it would be updated into the database.

VI SUMMARY

Humans, wittingly or unwittingly, are evolving their tools of trade to much a resemblance to themselves. The current steps taken might just model an aspect of human interactions. However, none could be said as to the future of human-machine interfaces as there would ultimately be a successor worthy of being called equal in cognitive strength with a human. Till then, it is suffice to be able to produce machines that could engage in complex and non-linear problems such as pattern matching and 3-D navigation and control.

Autopoiesis provides for a fresh perspective to the time-worn research of speech recognition. Thus, it is safe to say that adopting a self-organizing paradigm to speech recognition would open a door for the mainstreaming of autopoiesis into engineering and systems development. The importance of autopoiesis cannot be understated for there truly is a great need for a shift in the paradigms of research and development if there ever will be a promise of continued innovation and creation. Therefore, this paper is meant to investigate the potentials of autopoiesis rather than recreate another generic speech recognition model. After all, it is not speech recognition which is at stake, rather it is the study of cognition that is most affected.

REFERENCES

[1] Flanagan, James L. *Speech Analysis: Synthesis and Perception, 2nd Ed.* Springer-Verlag, Berlin. 1972.

[2] Markowitz, Judith A. *Using Speech Recognition.* Prentice Hall, New Jersey, pp. 26, 1996.

[3] Spina, Michelle S. and Zue, Victor W., "Automatic Transcription of General Audio Data: Effect of Environment Segmentation on Phonetic Recognition". In *Proceedings of Eurospeech 97*, pp. 1547-1550, Rhodes, Greece. Sept. 1997.

[4] Tebelskis, Joe. *Speech Recognition using Neural Networks.* PhD Thesis, Carnegie Mellon Uni., May 1995.

[5] Hunt, Melvin J. "Signal Representation". In *Survey of the State of the Art in Human Language Technology*, pp. 11 – 16, Nov. 1995.

[6] Varela, Francisco J. *Autopoiesis and a Biology of Intentionality.* CREA, CNRS – Ecole Polytechnique, Paris, France.

[7] Maturana, Humberto R. and Varela, Francisco J. *Autopoiesis and Cognition: The Realisation of the Living.* Reidel Publishing, London, 1980.

[8] Boden, Margaret A. "Autopoiesis and Life". In *Cognitive Science Quarterly (2000) Volume 1*, pp 117 – 145, June 1999.

[9] Maturana, Humberto R. and Varela, Francisco J. *The Tree of Knowledge: The Biological Roots of Human Understanding.* Shambala, Boston, 1992.

[10] Kohonen, T. "The Neural Phonetic Typewriter". In *IEEE Computer Volume 21*, pp 11-22, March 1988.

[11] Bose, N.K. and Liang, P. *Neural Network Fundamentals with Graphs, Algorithms, and Applications.* McGraw-Hill International Editions, 1996.

[12] Fausett, Laurene. *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications.* Prentice Hall, New Jersey, pp 169, 1994.