Cognitive Informatics: The Future of Spoken Language Processing ?

Roger K. Moore

Speech and Hearing Research Group, Dept. Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK
r.k.moore@dcs.shef.ac.uk

Abstract

There is no doubt that the past 50 years have seen spectacular progress in our scientific understanding of how human beings use and process spoken language, as well as in our technical ability to mimic such behaviour in practical computer-based systems. However, despite these impressive achievements, we still have a long way to go before these two areas of knowledge and discovery converge on a coherent 'theory' of spoken language processing; a theory that could serve to both explain the intricacies of human speech behaviour as well as support a truly ubiquitous technology for spoken language processing. This paper discusses the prospects for a deeper and more unified understanding of spoken language processing and concludes that future progress may come from the newly emerging field of 'Cognitive Informatics'.

1. Introduction

Spoken Language Processing (SLP) is an area of fundamental scientific importance. Not only is it arguably the most sophisticated behaviour of the most complex organism in the known universe [3][5], but automated SLP systems offer the potential to incorporate voice-based interaction into a wide variety of futuristic applications [12][13].

Indeed, over the past 50 years or so, our understanding of how human beings use and process spoken language has made immense strides, especially with the introduction of modern computer-based analysis tools and techniques. At the same time, our ability to create and implement practical systems for analysing and generating speech signals has grown from strength to strength, and what was once a commercial dream is now becoming a practical reality.

However, we are not at the end of the road – our level of understanding of spoken language is still quite modest in comparison to the sophisticated communicative behaviour exhibited by the average human being, and the full potential of truly ubiquitous spoken language technology cannot be realised using today's models and algorithms.

1.1. Spoken language processing by machines

There is no doubt that recent years have seen a substantial growth in the capabilities of spoken language technology, first in the research laboratory and more recently in the commercial marketplace. Progress has reached a point where large-vocabulary automatic speech recognition (ASR) is available for only a few tens of dollars in any high-street computer store, text-to-speech synthesis (TTS) can be heard delivering announcements at public bus stops and where the automated handling of conversational speech using a spoken language dialogue system (SLDS) is becoming a familiar feature for users of telephone-based Interactive Voice Response (IVR) systems.

These developments have come about *not* as a result of any deep insights into the way in which human beings process language, but largely as a consequence of the introduction of the data-driven/machine-learning approach to building spoken language systems (in which large corpora of annotated speech recordings are used to capture the variability of speech) coupled with the relentless increase in available computing power.

However, despite this acknowledged progress in constructing practical spoken language systems, it is nevertheless the case that the performance of such technology, either as individual components or as whole systems, still falls some way short of the capabilities of the human listener/speaker [35][41] particularly in terms of 'robustness' and 'flexibility' in real-world environments. This has been clearly illustrated in Lippmann's well known comparison of ASR and HSR (human speech recognition) accuracy [15] which revealed that ASR is at least an order-of-magnitude worse than that of human listeners across a range of tasks of varying difficulty.

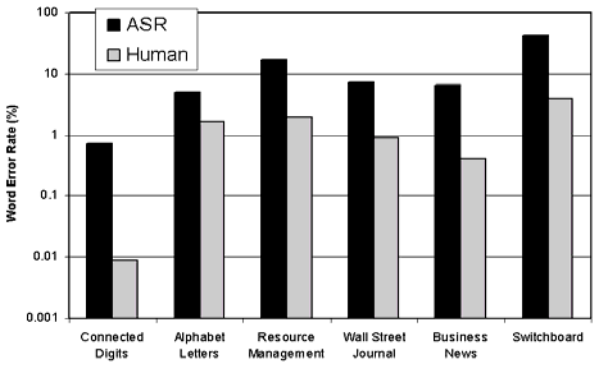

Figure 1: Comparison of human and automatic speech recognition performance on a range of different tasks (derived from Lippmann [15]).

This shortfall in performance of automatic spoken language systems has also been recognised by some of the industry's leading practitioners:

> *"The industry has yet to bridge the gap between what people want and what it can deliver. Reducing the ASR error rate remains the greatest challenge."*
>
> X. D. Huang [10]

"After sixty years of concentrated research and development in speech synthesis and text-to-speech (TTS), our gadgets, gizmos, executive toys and appliances still do not speak to us intelligently."

Caroline Henton [7]

Indeed it has been estimated that even if the present rate of incremental progress can be sustained, it will still take another twenty to forty years before ASR reaches the level of accuracy exhibited by a human listener [10]:

- transcription of read newspaper text by 2013
- recognition of alphabet letters by 2017
- transcription of freestyle speech by 2021
- recognition of digit strings by 2043

Unfortunately this represents the optimistic view; it has also been estimated that 100,000 to 1,000,000 hours of speech data would be needed to train such advanced ASR systems (see Fig.2) [26], and this must surely raise questions about the general approach. Similarly for TTS, it has been suggested that the size of database needed to capture 100 different talking styles and 10,000 different voices would be 5,000 Gbytes, and that our technology is not up to generating such databases automatically [11].

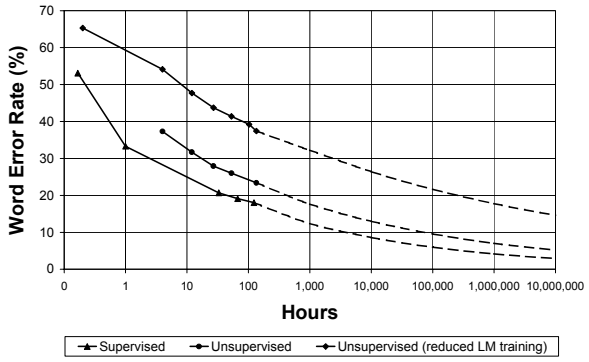

Figure 2: Extrapolated word error rates for increasing quantities of training data (taken from Moore [26]).

Another consequence of the current data-driven approach to spoken language technology is the fragility of the engineered solutions with respect to real user environments [2]. State-of-the-art ASR is quite poor at accepting input that is spontaneous, emotional, whispered, accented, disfluent, interrupted, contaminated, from the young/elderly/non-native, or which is rich in previously un-encountered words, expressions and behaviours. Similarly, although contemporary text-to-speech synthesis delivers output that is human-like and reasonably intelligible, it still lacks expressiveness and is non-reactive to the communicative situation [11]. Also, once trained, the behaviour of current spoken language technology systems is essentially fixed. Minor adaptation takes place but such systems do not learn to handle new concepts or tasks that arise in the course of an ongoing interaction [28].

Therefore, in order to move to the next generation of SLP systems, there is a growing perception that a paradigm shift is needed in the underlying algorithms and technology. Unfortunately, there is little consensus as to what particular direction to pursue, and many researchers believe that ASR will never equal the performance of HSR (see Fig.3).

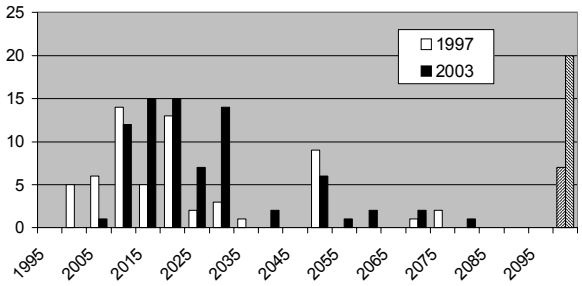

Figure 3: Responses to the question "When will speech recognition accuracy equal that of the average human transcriber." (taken from Moore [30]). The right-hand columns indicate 'never'.

1.2. Spoken language processing by humans

Human spoken language behaviour has been the subject of study for many years and a very wide spread of disciplines has become involved. A great deal is known about the physical properties of peripheral components such as the auditory and articulatory systems, but study becomes much more problematic at progressively higher levels of phonetic, linguistic and cognitive processing.

This lack of knowledge at higher levels arises because, unlike the engineers and computer scientists concerned with machine-based systems, researchers into human spoken language are faced with the challenge of understanding a system by mainly indirect (rather than direct) observation. A human SLP system cannot be taken apart, nor can its components be isolated in order to identify their functionality and contribution [25].

This situation has lead to the development of extremely sophisticated techniques for probing human spoken language behaviour, ranging from carefully designed acoustic tests of hearing, to experiments on the phonetic identification and discrimination of sounds, to psychological tests on the recognition and comprehension of words [17], to the latest neural-imaging investigations [33][40][44][45]. Insight has also been obtained from behavioural studies of human spoken language systems that have been damaged by trauma or illness.

Models relating to single aspects of human SLP exist [14]. However, because of the fragmentation of research across all the different levels of human SLP, there is little integration and models are often neutral with respect to the behaviour of models in other areas (for example, models of human word recognition are ambivalent about the exact form of the input to the word processing system, and phoneticians likewise do not commit themselves to claims about whether explicit phonemic identification is actually part of the human SLP system). Also, many models of human behaviour are descriptive rather than computational – an absolutely vital difference in respect of an attempt to determine how a system might actually work.

However, in recent times computational models of aspects of human SLP have begun to emerge (driven by the

early excitement from 'connectionism') [31], and this has given rise to a greater interest in integration, and in the relationship between human and machine behaviour. Indeed the most recent computational models of human word recognition draw heavily on the algorithms commonly used in ASR [32], and have now been extended to include real acoustic input [37][39]. These attempts to bridge the gap between ASR and HSR are very interesting and provide some useful convergence, but it seems to be leading to the unfortunate consequence that the new HSR models suffer from the same limitations as the automatic systems.

2. Future prospects

In order to progress our understanding of SLP, what is needed now is not to consider current machines as putative models of the human system [18] or humans as an enigmatic blueprint for future machines [23], but to look towards a 'unifying theory' that would be capable of explaining and predicting both human and machine SLP behaviour [27].

However, in order to do this it may be necessary to grow a new research community that is dedicated to this particular goal. This is not intended to be a divisive position, but a pragmatic suggestion based on the reality that the current research communities have their own drives and interests, viz. fathoming the design of parts of an extant system versus building and selling complete end-to-end systems [32].

Historically, there has been little overlap between the research communities working on SLP by humans and those working on SLP by machines. Indeed, an extremely wide range of disciplines has evolved each claiming part ownership of the problem, for example: acoustics, psychoacoustics, phonetics, linguistics, psycholinguistics, psychology, perception, production, cognitive neuroscience, brain imaging, human factors, signal processing, pattern recognition, computer science, machine learning, natural language processing, artificial intelligence, neurocomputing, engineering, graphics, virtual reality, interface agents etc. etc.

This growing dispersion and lack of coherence may now have an opportunity to be reversed thanks to the recent appearance of an entirely new discipline – 'Cognitive Informatics' (CI).

3. Cognitive informatics

Cognitive informatics is an emerging transdisciplinary field in the cognitive and information sciences that aims to forge links between a diverse range of disciplines spanning the natural and life sciences, informatics and computer science [48]. CI is founded on the conviction that many fundamental questions of human knowledge share a common basis - an understanding of the mechanisms of natural intelligence and the cognitive processes of the brain. CI views the brain as an information processing organ, and the new field aims to provide a coherent focus for *all* of the relevant areas of study, including:

- Autonomous computing
- Intellectual foundations of informatics
- Information models of the brain
- Informatics foundations of software engineering
- Expressive mathematics
- Internal information processing mechanisms
- Software agent systems
- Ergonomics
- Informatics laws of software
- Knowledge representation
- Neural computation

The appearance of CI and its community of like-minded researchers presents a unique opportunity for SLP to sit, not at the crossroads of many different disciplines (as it does now), but at the very heart of this new field (see Fig.4). Indeed if one subscribes to the statement introduced in Section 1 that *"SLP is the most sophisticated behaviour of the most complex organism in the known universe"*, then it can be argued that a 'unified theory of SLP' should constitute one of the core 'grand challenges' [9] in cognitive informatics.

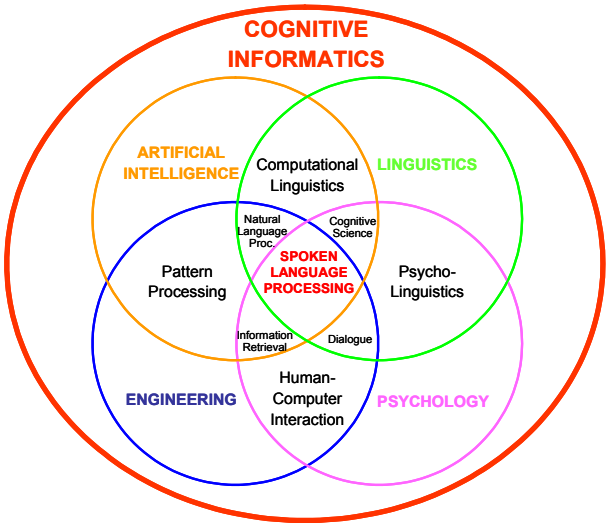

Figure 4: Spoken language processing at the heart of 'Cognitive Informatics'.

4. Towards a unified theory

What would such a unifying theory look like and how would it be different from what the community has right now [29]? Consider the core SLP behaviours illustrated in Fig. 5. The most obvious feature of this diagram is that it encompasses the entire 'speech chain' including speech perception and production, speech interpretation and expression and conversational interaction. Important aspects such as emotion and individuality are seen as relevant to both the interpretation of spoken input (i.e. determining who is talking and what state they are in) and to the expression of spoken output (i.e. overlaying personal and emotional behaviours). The figure also suggests the possible close-coupling of all these processes with a potential for unified internal representations and mechanisms that are ambiguous with respect to listening or speaking.

Fig. 5 also emphasises the multi-modal nature of SLP, i.e. an acknowledgement of the high degree of relevance of gesture, facial expression and lip movement as powerful accompaniments to the acoustic signal. What the figure is unable to capture is the essential process that underpins the behaviour of any spoken language system – the 'patterning':

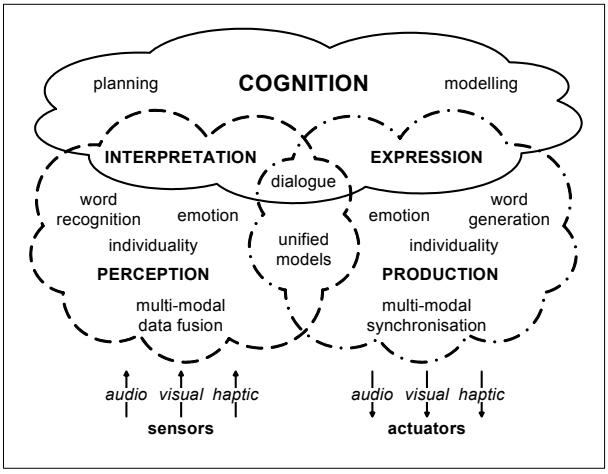

Figure 5: Illustration of the core behaviours in spoken language processing.

> *"Speech production and perception are essentially processes which relate a series of acoustic and visual events to corresponding cognitive activity. Speech mediates the expression and communication of ideas, concepts and information between different physical entities through a regularity of behaviour which is shared (and hence understood) by the participants. It is the regularity of behaviour - the patterning - which is the central object of study in all areas of speech research."*
>
> Roger Moore [20][24]

According to Moore [21], a theory of SLP should thus encompass:

- a-priori information about the regularities of speech (i.e. constraints in the form of descriptive knowledge about speech patterns, linguistic structures and the relationships between different levels of description, together with corpora of recorded speech/language material and their annotations)
- the representation of speech knowledge and information derived from actual speech data (i.e. the encoding of the constraints)
- the computation that must be executed in order to achieve the required transformations (i.e. algorithms for constraint satisfaction)

Of course a key issue is the nature of the encoding, i.e. the internal representations. What is required is a mathematical and scientific formalism for encoding a-priori information in a computationally useful form: a formalism which can (i) exploit regularities (patterns) in the data, (ii) generalise from seen data to unseen data (in order to exhibit robust and flexible behaviour in the face of changing circumstances) and (iii) use the minimum amount of information to achieve its goals in response to some efficiency criterion.

Current machine SLP approaches to encoding are limited in a number of areas. For example, techniques are based on theoretical assumptions that limit the ability to accommodate simultaneous asynchronous behaviour [42], dynamic patterning, decomposed dependencies [46][47] or minimal distinctions [19] – all acknowledged properties of spoken language. Similarly, only a small amount of research has been conducted into computational mechanisms that might underpin embodied and situated learning [37], including rapid-adaptation, long-term and high-level structural re-configuration and the process of language acquisition and/or evolution in a communicative environment [1].

Progress may also be highly dependent on the invention and/or discovery of computational mechanisms that reflect brain function at different levels of detail. However, at the present time there appears to no overarching theory of brain function that can be used as some form of design guide [34][4]. Useful insights can be gained from ideas such as episodic memory [43][8][16], the 'memory-prediction framework' proposed in [6] and 'Perceptual Control Theory' (PCT) [36], but real progress may come about through the growing body of work in cognitive informatics [49].

5. Summary and Conclusion

This keynote paper has argued that the future of spoken language processing lies in attempting to create a unified theory that would be capable of explaining and predicting both human and machine SLP behaviour. It is also been suggested that this can best be done by establishing such an objective as a grand challenge at the heart of the newly emerging field of Cognitive Informatics.

6. References

[1] Altmann G T. *The Ascent of Babel*, Oxford University Press, 1997.

[2] Cole R., Hirschman L., Atlas L., Beckman M., Biermann A., Bush M., Clements M., Cohen J., Garcia O., Hanson B., Hermansky H., Levinson S., McKeown K., Morgan N., Novick D., Ostendorf M., Oviatt S., Price P., Silverman H., Spitz J., Waibel A., Weinstein C., Zahorian S. and Zue V. "The challenge of spoken language systems: research directions for the nineties", *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 1-21, 1995.

[3] Dawkins, R. *The Blind Watchmaker*, Penguin, 1991.

[4] Fodor J. The Mind Doesn't Work That Way, *MIT Press*, 2001.

[5] Gopnik A, Meltzoff A N, Kuhl P K. *The Scientist in the Crib*, Perennial, 2001.

[6] Hawkins J. *On Intelligence*, Times Books, 2004.

[7] Henton C. "Fiction and Reality of TTS", *Speech Technology Magazine*, vol.7, no.1, Jan/Feb 2002.

[8] Hintzman D L. "Schema-Abstraction in a Multiple-Trace Memory Model", *Psychological Review*, 93: 411-427, 1986.

[9] Hoare T and Milner R. "Grand challenges for computing research", *The Computer Journal*, Vol. 48(1), pp. 49-52, 2005.

[10] Huang X D. *Making speech mainstream*, Microsoft Speech Technologies Group, 2002.

[11] Keller, E., "Towards Greater Naturalness: Future Directions of Research in Speech Synthesis", *Improvements in Speech Synthesis*, Keller, E., Bailly, G, Monaghan, A., Terken, J. and Huckvale, M. (eds.), Wiley & Sons, Chichester, UK, 2001.

[12] Kurzweil R, *The Age of Intelligent Machines*, MIT Press, 1990.

[13] Kurzweil R, *The Age of Spiritual Machines*, Phoenix Press, 1999.

[14] Lindblom, B., "Explaining Phonetic Variation: A Sketch of the H&H Theory", *Speech Production and Speech Modeling*, Hardcastle & Marchal (eds.), Kluwer, pp. 403-439, 1990.

[15] Lippmann R. "Speech recognition by machines and humans", *Speech Communication*, vol. 22, pp. 1-16, 1997.

[16] Maier V. and Moore R K. "An investigation into a simulation of episodic memory for automatic speech recognition", *Proc. InterSpeech*, 2005.

[17] McQueen J and Cutler A. "Spoken word access processes: an introduction", *Language and Cognitive Processes*, 16(5/6), pp. 469-490, 2001.

[18] Moore R K. "Speech recognition systems and theories of speech perception", *The Cognitive Representation of Speech*, Myers, Laver and Anderson (eds.), North Holland, pp 427-441, 1981.

[19] Moore R K, Russell M J and Tomlinson M J. "The discriminative network; a mechanism for focusing recognition in whole-word pattern matching", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Boston, 14-16 April, pp1041-1044, 1983.

[20] Moore R K. "Whither a theory of speech pattern processing?", *Proc. Eurospeech*, Berlin, 21-23 September, 1993.

[21] Moore R K. "Speech pattern processing: from blue sky ideas to a unified theory?", *Proc. Institute of Acoustics Conference on Speech and Hearing*, Windermere, November, 1994.

[22] Moore R K. "Pre-lexical processing: a view from ASR", invited lecture, *Workshop on Methods and Models of Spoken Word Recognition*, Max-Planc Institute, Nijmegen, 26-27 January, 1995.

[23] Moore R K. "Critique: The potential role of speech production models in automatic speech recognition", *Journal of the Acoustical Society of America*, Vol.99, No.3, pp 1710-1713, March, 1996.

[24] Moore R K. "Speech pattern processing", *Computational Models of Speech Pattern Processing*, K Ponting (ed.), NATO ASI Series F, Vol.169, Springer-Verlag, pp 1-9, 1998.

[25] Moore R K and Cutler A. "Constraints on theories of human vs. machine recognition of speech", *Proc. SPRAAC Workshop on Human Speech Recognition as Pattern Classification*, Max-Planck-Institute for Psycholinguistics, Nijmegen, 11-13 July, 2001.

[26] Moore R K. "A comparison of the data requirements of automatic speech recognition systems and human listeners", *Proc. Eurospeech*, Geneva, pp. 2582-2584, 1-4 September, 2003.

[27] Moore R K. "Towards a Theory of Speech Recognition: bridging ASR & HSR", Invited talk, *Workshop on Innovative Approaches Bridging Automatic and Human Speech Recognition*, University of Nijmegen, 17 November, 2003.

[28] Moore R K and Cunningham S C. "Plasticity in systems for automatic speech recognition: a review", *Proc. ISCA Workshop on Plasticity in Speech Perception*, London, June, 2005.

[29] Moore R K. "Towards a unified theory of spoken language processing", *Proc. 4th IEEE International Conference on Cognitive Informatics*, Irvine, CA, USA, 8-10 August 2005.

[30] Moore R K . "Results from a survey of attendees at ASRU 1997 and 2003", *Proc. INTERSPEECH 2005* Lisbon, 5-9 September 2005.

[31] Norris D. "Shortlist: a connectionist model of continuous speech recognition", *Cognition*, 52: pp. 189-234, 1994.

[32] Norris D, McQueen J and Cutler A. *Behavioral and Brain Sciences*, 23: pp. 299-370, 2000.

[33] Patterson R, Uppenkamp S, Johnsrude I and Griffiths T. *Neuron*, 36: pp. 767-776, 2002.

[34] Pinker S. *How The Mind Works*, Penguin Books, 1997.

[35] Pols L. "Flexible, robust, and efficient human speech processing versus present-day speech technology", *Proc. of the 14th Int. Congress of Phonetic Sciences (ICPhS-99)*, San Francisco, USA: pp. 9-16, 1999.

[36] Powers W T. *Behavior: The Control of Perception*, Hawthorne, NY: Aldine, 1973.

[37] Roy D and Pentland A. "Learning words from natural audio-visual input", *Proc. Int. Conf. on Spoken Language Processing*, pp. 1279-1282, 1998.

[38] Scharenborg O, ten Bosch L, Boves L. and Norris D. "Bridging automatic speech recognition and psycholinguistics: extending Shortlist to an end-to-end model of human speech recognition*", J. Acoustical Soc. of America*, Vol. 114(6), pp. 3023-3035, 2003.

[39] Scharenborg O, McQueen J, ten Bosch L and Norris D. "Modelling human speech recognition using automatic speech recognition paradigms in SpeM", *Proc. Eurospeech*, Geneva, pp. 2097-2100, 2003.

[40] Scott S, Blank S, Rosen S and Wise R. *Brain*, 123: pp. 2400-2406, 2000.

[41] Sroka J J and Braida L D. "Human and machine consonant recognition", *Speech Communication*, 45(4), pp. 401-424, 2005.

[42] Tomlinson M J, Russell M J, Moore R K, Buckland A P and Fawley M A. "Modelling asynchrony in speech using elementary single-signal decomposition", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp 1247-1250, Munich, 1997.

[43] Tulving E. "Episodic Memory: from Mind to Brain", *Annu. Rev. Psychol*. 53, 1–25, 2002.

[44] Tyler L, Russell R, Fadili, J and Moss H. *Brain*, 124: pp. 1619-1634, 2001.

[45] Tyler L, Bright P, Dick E and Stamatakis E. *Neuropsychologia*, 42: pp. 512-523, 2004.

[46] Varga A P and Moore R K. "Hidden Markov model decomposition of speech and noise", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Albuquerque, 3-6 April pp. 845-848, 1990.

[47] Varga A P and Moore R K. "Simultaneous recognition of concurrent speech signals using hidden Markov model decomposition", *Proc. Eurospeech*, Genova, September, 1991.

[48] Wang Y. "On cognitive informatics", *Brain and Mind*, Vol. 4, pp. 151-167, 2003.

[49] Zhang D and Kinsner W, *Proc. 4th IEEE Int. Conf. On Cognitive Informatics*, Irvine, CA, 8-10 August 2005.