ROMANIAN ACADEMY
Section for Science and Technology of Information,

UNIVERSITY "POLITEHNICA" OF BUCHAREST
Faculty of Electronics, Telecommunications and Information Technology

TECHNICAL UNIVERSITY OF CLUJ-NAPOCA
Faculty of Electronics and Telecommunications

Trends in Speech Technology

Proceedings of the 3rd Conference Speech Technology and Human-Computer-Dialogue "SpeD 2005"
Cluj-Napoca, May 13-14, 2005

Coordinator:
Corneliu BURILEANU

EDITURA ACADEMIEI ROMÂNE
Bucharest, 2005

CAN AUTOMATIC SPEECH RECOGNITION LEARN MORE FROM HUMAN SPEECH PERCEPTION?

Sorin DUŞAN, Lawrence R. RABINER

Center for Advanced Information Processing
Rutgers University
Piscataway, NJ 08854 U.S.A.
sdusan@caip.rutgers.edu, lrr@caip.rutgers.edu

Corresponding author: Sorin Duşan

Although a great deal of progress has been made during the last two decades in automatic speech recognition (ASR), the performance of these ASR systems, as measured by word recognition and concept understanding error rates, is still much worse than that achieved by humans, even for carefully read and articulated speech in quiet conditions. This performance gap (between machines and humans) increases even more in noisy conditions and for conversational speech. Steadily increasing computational speed and computer memory tend to impose fewer and fewer constraints on the types and the amount of recognition processing that can be brought to bear on a particular recognition task. In spite of the increased computation and memory, the state-of-the-art technology in automatic speech recognition appears to have reached a plateau in the past few years. New techniques and principles need to be invented or applied in order to substantially reduce the current performance gap in speech recognition between humans and machines. This paper presents some ideas intended to stimulate further research on applying knowledge and principles derived from studies of human speech perception to automatic speech recognition. Although the mechanisms of human speech perception (HSP) are not fully understood, some findings from neuroscience, physiology, cognitive science and psychology could potentially lead to new understanding and thereby stimulate the development of new techniques and architectures for automatic speech recognition that, eventually, will bridge and reduce the performance gap between machines and humans.

Key words: Automatic speech recognition; Speech perception; Auditory processing; Phonemes; Syllables; Words; Mental lexicon.

1. INTRODUCTION

Substantial progress has been made in the area of automatic speech recognition (ASR), especially during the last two decades, when techniques based on hidden Markov models (HMM) and artificial neural networks (ANN) were developed and constantly refined. At the beginning of the research cycle, when the rate of progress was high, many new ideas were proposed and evaluated, and the best of these ideas enabled greatly improved speech recognition performance (over previous generations of technology). One somewhat negative aspect of the feverish activity in devising new methods associated with HMM and ANN technologies was that there were very few researchers who were working on alternative recognition architectures and systems during this period. Hence, now that HMM and ANN technology has reached a state of maturity, it has been shown that the performance of such systems, in real working environments and with vocabularies and syntax that are required for real world applications, is significantly below that of human speech recognition performance. The error rates obtained by the best known ASR methods are often more than an order of magnitude higher than those of humans for 'clean' conditions. These differences in performance between ASR and humans become even larger in tasks that involve more realistic 'noise and background' conditions. Lippmann provides an excellent summary comparing and contrasting the performance of modern ASR systems with those of humans across a range of recognition tasks [1]. The current low rate of progress of modern ASR systems has generated renewed interest in the ASR community to think about developing new techniques and architectures for accelerating the rate of progress in bridging the gap between machine and human performance of recognition systems across a wide range of problems.

In order to gain an understanding of the new directions that are currently being pursued in speech recognition, it is worth reviewing the early history of speech recognition research. The first significant attempts to perform ASR, although primitive, were based on early understanding of hearing and human speech perception (HSP). These early speech recognition systems assumed that short speech segments corresponding to phonemes, dyads, syllables or even words could be uniquely mapped into the corresponding linguistic units by measuring the "spectral distance" between these segments (or some appropriate spectral representation of these segments) and a set of previously recorded and labelled templates. The resulting set of methods of speech recognition, based almost exclusively on simple pattern recognition methods and token exemplars called templates, formed the basis for speech recognition research for almost two decades, and was only recently displaced by the set of statistical pattern recognition methods as exemplified by HMMs and ANNs.

There is a long running debate regarding the inequality between ASR and HSP, both from the performance perspective, and from the involvement (or lack thereof) of the higher levels of cognition and understanding in the ASR processing. This debate leads to questions as to whether the two processes (ASR and HSP) could ever perform similarly so long as machines do not process speech like humans. One possible conclusion from this discussion is a strong lack of faith in ASR research and its ultimate ability to solve the speech recognition problem in realistic environments. Some have even questioned whether it is worth continuing and supporting such research [2]. Current ASR techniques (like HMM and ANN) are data-driven. Such data-driven models infer or learn the relevant speech structures from large quantities of training data and use relatively simple speech models to map acoustics to (context-dependent) phones and words. These data-driven techniques and the resulting models are often cited for incorporating too little (linguistic and acoustic) knowledge about speech and about the processes of speech perception in humans. While a large number of researchers from the ASR community believe that new directions and discoveries are necessary to match or exceed human performance, others, often a more conservative group, think that having significantly more training data will be sufficient to achieve this goal. Government organizations are, in principle, open to such new directions [3], although they are reluctant to completely abandon the mainstream direction in which they previously invested considerable resources. This leads to a third school of thought, namely that it is possible to maintain and build on existing technology by supplementing it with appropriate linguistic and acoustic information, thereby significantly improving performance without having to completely duplicate existing human recognition capabilities [4].

While using new knowledge from HSP to develop new methods in ASR and improve recognition performance appears to make sense, there are a number of practical problems and difficulties associated with this course of action. First, there is no complete understanding of the mechanisms and processes that take place in human speech perception and speech comprehension; hence we are always working with an incomplete and inaccurate representation of linguistic and acoustic knowledge of speech. Second, not every complete or partial discovery in speech perception will lead to an improved computational model for ASR; many times new linguistic or acoustic knowledge cannot be easily implemented in the current state-of-the-art technologies. Third, the area of speech perception and understanding is a multidisciplinary venture where discoveries and theories emerge from various fields such as neuroscience, psychology, linguistics, cognitive science, etc. These fields are usually not well monitored and understood by engineers and computer scientists who implement new technologies. Finally, understanding how the brain works poses formidable difficulties even for simpler, more specialized functions such as hearing. The auditory cortex contains some 100 million neurons [5], whereas the total number of neurons in the human cerebral cortex is about 10 billions and in the brain is about 100 billions [6]. The enormous number of neural cells in any specialized cortical brain function is just one complication in trying to explain psychophysical phenomena. A second difficulty is the complicated interconnections among these neurons and also between them and the thalamus and other cortical areas of the brain (where there are 60 trillion synapses in the cerebral cortex according to [6]). A third difficulty is the microscopic nature of these neurons and their interconnections. These difficulties often make the study of speech perception rely on external observations and behaviour that only indirectly (and often nonlinearly) represents specific observed phenomena. Thus, most of the more biologically-inspired auditory models, proposed for use in ASR systems, account primarily for the processes that take place in the cochlea and at the lower levels of the neural transduction process in the auditory pathway. Among such models are the Lyon Cochlear model [7], the Seneff joint synchrony/mean-rate model [8], and the Hermansky Perceptual Linear Predictive (PLP) model [9].

The goal of this paper is to emphasize new understanding and discoveries emerging from the set of multidisciplinary areas related to speech perception and to encourage the integration and the implementation of this new knowledge into new and existing technologies for ASR. Although the processing and integration of this new knowledge from so many diverse sources is not trivial, it is precisely what the brain does well, i.e., integrating information from lower levels to build higher levels of meaning and cognition. Emphasizing the role of discovering new knowledge by integration does not mean that the essential role of scientific discoveries in any specific field or area should be de-emphasized or, that any integration of information results in higher levels of knowledge and understanding. Thus this paper compares some general characteristics of HSP with techniques used in ASR and discusses some issues that could be responsible for the performance limitations of current ASR technologies. It is not yet known what type of technical solutions would overcome these problems, or even if the solutions to these problems would eliminate or significantly reduce the gap between HSP and ASR. However, this paper encourages researchers to explore such ideas both theoretically and experimentally. It is anticipated that future advances in the understanding of the brain processes together with new technological advances in computing will lead to new technologies that could eventually eliminate the gap in performance between ASR and HSP.

2. HOW WORDS AND CONCEPTS EMERGE AND ARE RELATED IN THE BRAIN

The sensory systems serve the role of acquiring and transmitting specific information (e.g., auditory, visual, tactile, etc.) to the brain or central nervous system (CNS). The brain's ability to build cognitive codes from the sensory information systems is realised by exposure and learning and to some extent it is hereditary. However, the capacity of the brain to encode (e.g., learn) and decode (e.g., recognize or recall) specific sensory information is not limited to speech and existed long before the spoken language emerged [10]. This capacity served a vital role in the survival of the human race, e.g., by recognizing food or predators in the environment. However, these primitive visual, auditory, tactile or olfactory codes corresponded to existing entities in the environment (e.g., water, trees, reptiles) and were not 'intentionally' created by human beings for the purpose of encoding a communication message. Here 'intentionally' refers to the capacity of an individual to produce or manifest, at one specific time, one of a number of distinct states (messages) for the purpose of communicating that state to other individuals. The brain has the capacity to decode, at a specific time, a static message (one state out of a finite number of states) and also to decode temporally encoded messages which can be defined as a temporal sequence of different states that, as a whole, represents a new message (a unique combination of states). This latter characteristic naturally involves the use of memory. Speech serves as a means of communication by temporally encoding messages. Speech is certainly not the first communication code ever invented. However, it is probably the most important one in the evolution of humans. Some authors even attribute the rapid intellectual evolution of our specie to the emergence of speech and language.

The brain can store and recognize an enormous number of codes representing concepts that are learned and built through exposure of the individual to the environment. However, the brain has a finite capacity for storing these codes or concepts due to the finite number, although very large, of neural cells and synapses. The emergence of spoken language enabled the brain to associate perceptual concepts to linguistic forms or units perceived as a temporally encoded message through the variation in time of the acoustic properties of the emitted sounds. The language uses discrete entities such as morphemes, words, phrases and sentences to represent concepts or meanings. The number of codes or concepts that can be perceived by the brain is believed to be higher than the number of words in a language [11]. There are indeed many native concepts that do not have corresponding linguistic forms (morphemes, words, or phrases) in a language and thus cannot uniquely be described linguistically. In general, the relation between words and concepts is not a one-to-one mapping [12]. However the number of concepts or meanings that can derive from language can theoretically grow infinitely due to the endless possible combinations of the linguistic forms. This certainly does not happen in the brain and an infinite number of theoretical combinations of linguistic forms do not have any associated meanings or concepts in the brain. Morphemes, words and, to some extent phrases, can be associated with primitive concepts that represent recognizable codes built and stored in the brain's memory. It is believed that human language processing uses three types of interacting structures, each stored separately in the brain: a) the concepts structure which includes all non-linguistic perceptions, actions and representations; b) the linguistic structure which includes the mental lexicon and the set of syntactic rules; and c) the mediation structure which interconnects the first two structures [13]. The first neural structure is located in both cerebral hemispheres, whereas the other two are mostly located in the left hemisphere. The word represents a special code perceived and learned through the auditory sensory system, and only later in life is associated with reading and writing. Its role is to activate other concepts or codes in the brain that in general are not perceived and learned through hearing and thus serves the purpose of communication among individuals. However, the emergence and use of language helps humans not only to communicate but also to build in their mind a structural representation (with complex levels and interconnections) of everything they know about the world, including themselves. Although the word-concept mapping is not one-to-one, it can be considered that the mental lexicon is used to build linguistic representations that duplicate somehow the mental perception and knowledge of the world. Even though the words and the concepts are starting to emerge in the brain of an infant during the first year of life, the acquisition of meanings (association between words and concepts) only starts during the second year of the child's life [14]. The codes of the mental lexicon can naturally be activated through perception and recollection or through thinking. They are stored in the long time memory (LTM) of the brain. Long phrases and sentences might not be stored as unique perceptible codes but rather as sequences in LTM of more elementary codes.

The activation of the codes in the mental lexicon induces the activation of some non-linguistic codes in the brain. Conversely, the activation of non-linguistic codes can induce a pre-activation of associated elements in the mental lexicon. These activations last for some interval of time so the perception of a word code or of a non-linguistic concept can be influenced by the previously activated codes or perceptions. When perceiving a linguistic message, this phenomenon, known as a contextual effect, influences the perception of the current word by the previously perceived words, phrases and sentences. It is expected that in speech perception both the word codes and the corresponding non-linguistic codes (concepts) become sequentially activated. Their activations likely overlap for some time. However, these mappings or interactions between the mental lexicon and the non-linguistic repository of concepts cannot naturally be shut down. Such a permanent interaction can be better understood by analysing the McGurk effect in speech perception which shows, indirectly, that conscious interruption of a neural channel of perception is practically impossible [15]. It appears that the perceptual information and the contextual information make joint contributions to the perception of words and these contributions are independent and processed in a similar way [16]. It is thus worth questioning the legitimacy of the expression 'human speech recognition' when humans listen to speech in a known language. Such questions can be appropriate when listening to nonsense speech, which does not happen so often, or to speech in an unknown language. The difference in performance between the two cases would reflect the contribution of the linguistic (lexical, syntactic, semantic) and non-linguistic layers of codes representing concepts to the perception of speech sounds. In ASR, however, this phenomenon is substituted simplistically by a single top-down processing scheme imposed by the syntactic rules and by word probabilities, which also takes place in speech perception. According to this simple paradigm, only words following the structure of the pre-programmed language rules (the system syntax) can be recognized and the recognition of a specific word (from a fixed lexicon) is influenced by a few preceding words (the n-gram language model estimated from a lexical corpus of training data). Thus, context effects are only accounted for in ASR through pre-programmed language rules and through word combination probabilities that are usually inferred (or learned) from the training data. In current ASR technologies there is no direct process that implements the local influence of the semantic conceptual codes (as opposed to the syntactic context) upon the recognition of current and future words in a sentence. Instead, this process is simplistically simulated by syntax and word probabilities derived from linguistic regularities.

3. SPEECH CODE AND INFORMATION PROCESSING

The speech signal represents a temporally encoded message. Speech employs the time dimension for the purpose of encoding discrete messages or codes. The speaker encodes in a non-stationary acoustic signal (generated through complex sound interactions in the vocal and nasal cavities) a certain linguistic form or

structure (e.g., a word, a phrase or a sentence) for the purpose of communicating or producing a specific message (or concept). The listener decodes the linguistic message embedded in the acoustic signal as a variation in time of the acoustic properties of the signal. The structure of the speech code reflects the structure of the language. These structures reflect the constraints imposed by anatomical, physiological and efficiency factors. Through evolution these structures and the speech code became highly optimised in order to enable highly efficient communication among individuals.

The codes activated in the brain can be perceived from a static source (e.g., a picture of an apple) or from a time-varying source (e.g., a speech segment containing the word 'apple'). In the former case probably only the conceptual (semantic) code is activated to reach consciousness, whereas the word code, although pre-activated, may or may not reach this level. In the latter case it is likely that the word code is activated first and then the conceptual code reaches the activation, but this happens so fast that at a conscious level they appear as a single event and not as two separate events in time. In either case the brain employs short time memory (STM) during perception.

It is known that the capacity of the brain to perceive, recall or activate (by thinking) linguistic or the non-linguistic codes in a unit of time is very limited [17], although the number of such learned codes stored in the brain is large. The brain works as a parallel processor of information provided by different sensory systems. The channel capacity of the brain to perceive information from a single sensory system along a single physical dimension is estimated to be at most 3 bits (8 different states or alternatives) [17]. This channel capacity increases by employing multiple simultaneous dimensions or modalities but it does not represent the summation of those corresponding to individual dimensions or to different modalities (sensory systems). The total channel capacity of the human brain to perceive parallel information from various sensory systems it is still unknown [18]. This total channel capacity is also estimated to be subject to a great deal of variability across individuals [18].

The speech code carries a certain amount of information which can be more easily evaluated at a word level or at sub-word levels as represented by phonemes or syllables. The average amount of information, measured in bits, generated by the event of selecting a symbol i from a group of N distinct symbols can be described by the entropy H as follows

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i, \tag{1}$$

where p_i represents the probability of occurrence of symbol i. This amount of information can be calculated in the case of selection or activation of one word i from the total number of words N stored in the mental lexicon. It is estimated that on average a 6 year old child possesses a lexicon of about 14,000 words whereas a high-school graduate has about 45,000 words in his lexicon [19]. If all words are considered equally probable then the entropy of one word event is 13.77 bits in the case of a 6 year old child and 15.46 bits in the case of a high school graduate. In reality these entropy values are lower because words occurring in speech do not have equal probability. A similar calculation of the average information can be applied to phonemes or syllables. If, for example, there are on the order of 42 phonemes in English, the average information per phoneme is 5.4 bits when the phonemes are considered equally probable and 4.7 bits when observed phoneme frequencies are used [20]. Since the number of syllables in English is much higher than the number of phonemes, but lower than the number of words, the average information per syllable would be comparable to the average information per word.

The capacity of the brain to process speech information, along with other anatomical and physiological constraints, should be reflected in the temporal structure of the speech signal. This means that the rate of speech events per second could not be higher than the rate at which the brain can process information, or otherwise the information would not be correctly perceived. Without discussing the role of phonemes in speech perception, if any, the information rate of conversational speech (in which phonemes occur at a rate of about 10 per second) is about 50 bits/s [20]. The analysis of conversational speech reveals a high optimization in the use of words because the majority of the most frequent words are monosyllabic [21]. These data reflect a correlation between the frequency of the words and their duration --- short words occurring more frequently than long words. In conversational speech, at a rate of 3 or 4 words/s [21], the average information rate at the word level would be between 46 bits/s and 61 bits/s for a lexicon of 45,000 words (assuming they have equal probability of occurrence and less if the word frequencies are employed). This shows that the rate at which the brain processes the speech information would be roughly the same at the phoneme level and at the word level. When other sensory channels or modalities are employed simultaneously with speech, the capacity of the brain to process all the information approaches a plateau and pushing more information above this limit results in some loss of information [18]. The fact that the human brain receives speech information at approximately the same rate, irrelevant of whether the speech unit is either the phoneme or the word, seems to support the design principles in current ASR technology. However, most of the current ASR approaches employ phonetic models (e.g., monophones or triphones) and not word models [22], although this is motivated by computational factors and not information processing rates. Hence a key question going forward for ASR systems is whether a more detailed theory of information processing in the brain would help the quest for the fundamental units of speech perception, or would it make it even more difficult.

4. UNITS AND MODELS OF SPEECH PERCEPTION

Computational models of speech perception are heavily dependent upon the choice of the units that are considered the building blocks of speech. It is not known if the complex temporal acoustic stimulus encoded by speech activates the word code directly or through a pre-process in which sub-word units, such as syllables, diphones, phonemes or features, are activated first and then some type of combining process activates the word-form. This probably represents the most intriguing question in speech perception [23], and there is a large body of studies and publications seeking answers to this question. Extensive reviews and discussions on this issue can be found in [24] and [25]. A definitive answer to this question could have important consequences for the implementation of ASR engines.

If the fundamental unit of speech perception were one of the sub-word phonological units (syllables, diphones, phonemes or features) this could be justified by some economic principles. Since words share all these sub-word forms, an economic structure in speech perception could be beneficial. The discussion here considers the case of the phoneme being the fundamental unit in speech perception, but many principles derived below would be similar when using other sub-words units of speech as the fundamental unit. An economic processing structure can be represented by a phonetic tree in which nodes correspond to combinations of phonemes and branches represent phonemes. All words that share the same combination of the initial phonemes up to a certain position would be processed by activating the same path through the nodes and branches up to that phonetic position. This phonetic tree would represent the whole mental lexicon as well as the phonotactic constraints of the language (the fact that not all the combinations of successive phonemes are possible in a language). This would require less storage memory and less computation than when each word is stored and activated using a separate memory for each of its phonemes. Due to its memory and computational efficiency this theory was adopted by many supporters in HSP and was efficiently implemented in ASR technologies. The brain possesses an enormous parallel computational power and it is not known if such an economic principle could have affected and modelled the processes and architectures of speech perception. However, in spite of its efficiency characteristics, this theory cannot completely account for some phenomena. For example, if humans employ such a phonetic tree in speech perception, how would this theory explain the fact that listeners are able to perceive a string of phonemes representing a nonsense word, since the phonetic tree does not contain such words. Or, if the phonetic tree represented all such possible words, how could the theory explain the learning of these neural codes corresponding to these words since most of them were probably never encountered during the learning process of the language.

Some theories support the idea that either the phonological features or the phonemes are the fundamental units of speech perception and all other phonological forms are built by successive combinations of these units, although they do not specifically support the idea of employing an efficient phonetic tree in speech perception. The acoustic invariant characteristics of these units could support these theories but these acoustic invariant properties are difficult to identify and several investigators have spent decades trying to find such acoustic invariant characteristics of the phonemes of English [26], [27]. A different view supports the idea that, for example, for place of articulation of the stop consonants, the invariant cues are not static but dynamic (spread across a phonological sequence) [28], which means that they cannot be associated with a single distinct feature or phoneme. A recent study showed that the spectral variability at the center position of phonemes is higher than that at the transition position between successive phonemes, and this difference is statistically significant [29]. Could this be an argument against the phoneme as the fundamental unit in speech perception?

Many studies of phoneme monitoring based on measuring human reaction time (RT) showed that the time to identify phonemes in syllables is higher than the time to identify the syllables [30]. However, using RT in the search for the fundamental units of speech perception could be deceiving since it can reflect some other processes that take place in the brain whose phenomena and epi-phenomena can influence the results. The conclusions from a large body of studies suggest that the identification of syllables and words precedes the identification of phonemes and thus that phonemes in speech are not perceived but inferred from the perceived words and syllables [25]. There is also strong evidence that illiterate adults cannot, or have great difficulties, segmenting speech phonetically. It is suggested that the ability to segment phonetically is acquired through learning how to read and write in alphabetical languages. This is supported by evidence that in languages using non-alphabetic writing systems, such as Chinese, listeners non-familiarized with any alphabetic writing system of other languages cannot segment phonetically.

Some studies of disorders in speech production support the role of the phoneme in speech processing. The idea is that when speakers make mistakes and replace a phoneme in a word with another phoneme which leads to a nonsense word, they use the phoneme as the fundamental unit in speech. However, even if the phoneme played a fundamental role in speech production, this does not mean that it would play the same role in speech perception. Up to here the issues discussed in this paper were not focused on and did not dwell on speech production since this paper is concerned with speech perception and automatic recognition of speech. However, it should be mentioned that an influential theory of speech perception is rooted in speech production. This theory, called the "motor theory of speech perception", presupposes that the listener perceives speech through a specialized 'module' in the brain that make direct use of the patterns and processes involved by the speaker during speech production [31]. This theory is strongly criticized however by the proponents of the auditory theories of speech perception who consider that perception does not involve any production processes. The supporters of the motor theory then advocated the idea that 'speech-is-special' because it involves two related processes: production and perception. It is no doubt that the two processes are related, but it is not clear why one of them could not function without the other? Totally deaf persons are capable of producing quite intelligible speech. Their only perceptions of speech are visual and proprioceptive (for their own speech). The auditory theories emphasize the role of auditory patterns in the perception of speech, and thus the independence of the speech perception process from the speech production process. These theories consider that the acoustic patterns of speech units are processed in a bottom-up manner and they do not rely on any motor process to activate the conscious linguistic perceptions. An example of such an auditory theory of speech perception is presented in [32], whereas a pure pattern recognition (bottom-up) approach to speech perception is presented in [33].

Connectionist models of speech perception employ excitatory and inhibitory interactions among various units which lead to the perception of words. In one such connectionist model, called TRACE [34], the activation of features produces the activation of phonemes, which produces the activation of words. The features are considered some fundamental phonological characteristics of speech (e.g., voiced, nasal, anterior) whose combination leads to specific sounds or phonemes. There is, nevertheless, criticism of this type of model. Why should listeners indirectly employ the recognition of some 'abstract' features whose properties are not acoustic but articulatory (related to the positions and functions of the articulators) and then combine these features to recognize phonemes and then words?

Other studies suggest that the fundamental units of speech perception are the syllables. In such a model, called the Fuzzy Logic Model of Perception (FLMP), the fundamental units of speech perception are believed to be the CV syllables, the VC syllables and the vowels, where V is a vowel and C is a consonant or a consonant cluster [35].

Due to a large number of studies providing various and sometime contradictory results and conclusions some researchers suggest that probably no fundamental unit of speech perception exists and the perception take place through the interactions of various levels of representation [36]. A similar approach to the fundamental units of speech perception, based on the Adaptive Resonance Theory (ART) [10], considers that the question about the fundamental unit in speech perception is misguided and it is an ill-posed problem [23].

According to this view, the units of speech perception are those features or phonological units that reach resonance at a perceptual conscious level. In other words, the perceptual units in speech are what the consciousness (or attention) focuses on. However, this type of view does not answer the question whether the perception of spoken words involves a pre-processing stage (either conscious or not conscious) in which sub-word units such as phonemes or syllables are perceived first and then concatenated to lead to the perception of words. The hypothesis that there is no unique, or fundamental, unit in speech perception and that this unit depends upon the focus of attention of the brain does not exclude either the possibility of serial processing (concatenation), or the simultaneous involvement of multiple types of phonological units. Since in ASR the most successful methods use the concatenation principle, it would be extremely important to know if the same process takes place in the human brain during the perception of speech.

Extensive evidence against the phoneme as the fundamental unit in speech perception makes more and more researchers believe that the syllable or word would be a better candidate for this unit [24], [25], [37]. The search for the fundamental units of speech perception is still underway because it is considered one of the most important problems in speech perception and still does not have a complete answer. Nevertheless, the search also continues for better models of speech perception. This search will probably continue until a better understanding of the underlying process emerges along with a more accurate model that can explain all the important phenomena observed in speech.

5. A COMPARISON BETWEEN HSP AND ASR

All ASR approaches are more or less inspired by principles from speech perception. A detailed comparison between HSP and all ASR methods is not possible since many aspects of HSP processes are not understood, but such a comparison can be made at some general level. Although there are various types of ASR techniques, they have many common processing stages and characteristics. In this section the comparison between HSP and ASR is made only along the following dimensions: a) architecture and levels of organization; b) spectral analysis and feature representation; c) top-down information processing; d) speech units; e) speech segmentation; and f) coping with speech variability.

5.1. Architecture and Levels of Organization

The most important architectural difference between HSP and ASR is that the former involves a large parallel neural processing system whereas the latter, based on computers, uses a serial processing system. Functionally, the former uses millions of neurons whose information processing rates are relatively low (a neuron can fire at a rate of approximately less than one thousand times per second) whereas the latter usually employs one microprocessor whose processing rate is very high (a microprocessor can currently work at a rate of about one billion instructions per second). It appears that these differences can be, to some extent, compensated for in ASR by the high processing rate of the microprocessors.

Another distinction is represented by the higher number of levels of organization in HSP (from the auditory system) than in ASR. In humans the spectral-temporal information from approximately 3,500 inner hair cells (IHCs) and 12,000 outer hair cells (OHCs) along the basilar membrane is transmitted by approximately 30,000 afferent fibres (90-95% receiving from the IHCs [38]) in each of the auditory nerves to approximately 90,000 neurons in the cochlear nucleus. Additional processing occurs at higher levels using the 34,000 neurons in the superior olivary complex and trapezoidal body, the 38,000 neurons in the lateral lemniscus, the 400,000 neurons in the inferior colliculus, the 500,000 neurons in the medial geniculate body and the 100,000,000 neurons in the auditory cortex [5]. Another hierarchical organization is represented by the six layers found in the auditory cortex. Important interconnections are also involved among these levels. It is believed that the auditory pathway below the cortical level plays a much more important role than simply transmitting the information to the higher processing levels. Such complex interconnections and the increasing number of carriers of information found in the auditory pathway are not usually implemented in existing ASR systems. Some ASR methods (based on ANN) employ multiple hidden layers of processors, usually 2 to 5, and time-delayed inputs to allow the network to capture and model nonlinear mappings and temporal correlations [39]. However, these architectures are quite homogeneous and it is unlikely that the number of levels and interconnections accurately reflect those employed in HSP.

An important distinction between ASR and HSP comes from the existence in humans of various parallel arrangements in the thalamocortical auditory pathway, apparently specialized to transmit and process distinctive properties of the sensory information [38]. At least three principal parallel pathways were found that correspond to a tonotopic system, a non-tonotopic system and a polysensory (multimodal) system. However, the exact contribution of these parallel systems to HSP is currently unknown. The specialization of populations of neurons was also found in the visual system where different groups of neurons process different attributes of images, such as form, colour and motion [40]. In addition other groups of neurons appear to be specialized to represent more detailed characteristics of the acoustic signal. Neurons are specialized to have the highest sensitivity at a specific characteristic frequency (CF), or at a specific threshold (TH) of the sound pressure level, or to have a specific spontaneous activation rate (SA), firing range (FR), or dynamic range (DR). These types of specializations in processing the acoustic properties are not found in ASR, where the acoustic features are homogeneous, although they do represent the frequency scale in a non-linear manner.

Another architectural distinction is represented by the redundancy offered by large groups of neurons in transmitting the same or similar information to the higher levels of processing. In the cochlea each inner hair cell transmits the spectral information to 10 or more fibres in the auditory nerve. The same type of redundancy is found at all the upper levels of processing, although the distribution and combination of information could be much higher since a typical neuron has between 1,000 and 10,000 synapses. In the brain the sounds are processed by a continuously increasing number of neurons (many carrying redundant information), whereas in ASR the sounds are parsimoniously represented by a reduced number of features (from hundreds of speech signal samples to tens of spectral features). It appears that the principle of parsimony plays a much higher role in ASR than in HSP, whereas the principle of redundancy plays a much higher role in HSP than in ASR (if any).

5.2. Spectral Analysis and Feature Representation

In humans the spectral analysis is performed along the basilar membrane of the cochlea by some 3,500 IHCs and 12,000 OHCs. Although it is believed that only the IHCs transmit 'important' ascending information to the higher auditory levels, it appears that OHCs contribute significantly to the frequency selectivity and sensitivity of IHCs. The approximate 30,000 neural fibres in each auditory nerve represent the acoustic signals by a myriad of firing rate patterns derived from all these neurons. Each of these neurons responds only to a specific frequency range and has specific characteristics (CF, SA rate, TH, DR and FR). Hence, there is a high specialization among these neurons in order to represent the entire frequency and dynamic range of human audition. In ASR the spectral analysis of the acoustic signal is performed usually at a few hundred frequency points (e.g., 512 Fourier magnitudes) which are then reduced to 10-20 spectral dimensions or features, usually with no correlation (or redundancy) among them (e.g., Mel Frequency Cepstral Coefficients - MFCC). Although both representations are time-dependent (time varying), in HSP the acoustic features are represented by firing rate (frequency) patterns and in ASR they are represented by magnitude patterns. The HSP representation is in a high-dimensional space (30,000 and continuously higher in the higher levels), whereas the one in ASR is in a low-dimensional space (10-20). Here, the so-called dynamic features in ASR are not counted because they represent a technical solution to a more complicated processing stage that is performed in the higher levels of the auditory pathway. A high-dimensional heterogeneous representation in HSP is replaced by a low-dimensional, rather homogeneous, representation in ASR.

5.3. Top-Down Information Processing

In HSP, and hearing in general, there are a large number of top-down connections represented by the efferent fibres at all the levels of the auditory system. Although their functions are far from well understood, it is believed that they play an important role in audition, including HSP. Various studies based on animals showed that the recognition of vowels in identification experiments was seriously degraded when efferent fibres in the auditory nerves were cut. It appears that these types of top-down interactions exist at all hierarchical levels in hearing, especially in the layers of the auditory cortex where syntactic and semantic information is combined with bottom-up information. In ASR such top-down connections are only implemented at high levels by emulating semantic and syntactic constraints with linguistic rules and word probabilities (n-grams). The efferent connections in the auditory pathway do not usually have any direct equivalent at the lower levels of information processing in ASR (e.g., features are not affected by top-down information). This could be a reason for the lack of robustness of the ASR features.

A related but distinct dissimilarity between HSP and ASR is represented by the existence of another major level of the architecture and process in HSP, represented by the multimodal concept code level, in addition to the word code level and the syntactic-semantic level that are also implemented in ASR. The syntactic-semantic level influences the word recognition in both systems but the multimodal concept code level is either not existent in ASR or is represented in a one-to-one manner by the word code level. This is not a minor distinction since in HSP the concept-word interaction is bi-directional and the concept code level is grounded in multiple modalities (e.g., sensory, motor, somatic). In ASR the word selected by recognition is only influenced by the bottom-up (acoustic) and the top-down linguistic constraints (syntactic and semantic) that are also present in HSP. In ASR applications such as command and control applications, a relatively small number of possible sentences are allowed to be recognized, and consequently syntactic-semantic constraints are indeed imposed upon the recognition, but these constraints are only implemented by the whole sentences and not by individual concepts as represented by the multimodal concept code layer in HSP. Individual word-concept interactions are simulated in ASR only at a linguistic level by using word probabilities (e.g., bi-grams, tri-grams, etc.) but this is a rather simplistic process and does not come from the multimodal environment context but from previous utterances (words). In ASR the recognition of the words of a sentence such as ‘Close the red valve’ can be influenced by top-down information by imposing syntactic and semantic constraints as well as word probabilities, but the perception of these words in HSP is also affected by many other multimodal concept codes which are actively involved in the context (e.g., ‘boss’, ‘pipe’, ‘job’, ‘wrench’, ‘water’, etc.), not only by the words (and their meanings) of the previous utterances.

5.4. Speech Units

All the important speaker-independent ASR approaches use the concatenation principle to represent words by successive phonemes. The fundamental unit of processing is thus the phoneme, which is usually represented by a context-dependent model (e.g., triphone, demiphone). Words are represented in the pronunciation lexicon as concatenations of phonemes, similarly as they are represented in writing by a concatenation of letters. However, in HSP, as discussed above, it is unlikely that the phoneme and the concatenation principle play the only central roles as in ASR. A variety of experimental and theoretical studies provide more and more evidence that in HSP some holistic processes employing words or syllables are likely to also be involved. Such approaches in ASR currently would probably cause computational problems and training problems due to insufficient speech data.

5.5. Speech Segmentation

Current ASR techniques, such as the HMM recognizer, perform the search for the best sentence on whole sentences by combining the acoustic and linguistic information and searching a lattice of words and phones for the best hypothesis. The recogniser exploits pauses or silence intervals in the incoming speech for segmenting the utterance into sentences (or phrases). Because these sentences usually obey grammatical rules, these rules can be imposed by top-down syntactic-semantic constraints, and thus the recogniser performs better when the recognition and final segmentation are performed on the whole sentence. However, sentences can last for a few seconds or more and it is know that humans do not usually wait for the end of the sentence in order to recognize words or phrases within the sentence. In ASR the segmentation of the input utterance into individual words (various hypotheses) leads to computing confidence scores and selecting the best hypothesis for the final recognition. The segmentation into words precedes the recognition and there is usually no local effect of the recognition of the current word on the segmentation of the next word. Of course the final recognition retrieves back the corresponding (best) word segmentation but this happens for the whole sentence.

In HSP it appears that the perception of the ‘current’ word plays a much more important role on the identification of the onset of the following word, whose identification in turn, plays a very important role on the recognition of that word. This might be also influenced by the fact, discussed in Section 5.3, that in HSP the meaning of an individual word usually is perceived at run-time (‘instantly’) and not after the completion of the whole sentence, although there are situations when that happens (sometime the recognition of the meaning of a current word depends upon a word or phrase that only comes after a few more words). It should not be understood that the ‘current’ word is processed without a certain delay or that this delay is always constant regardless of the identity of the word and the content of the sentence. In fact, some words (in particular long words) appear to be recognized before they ended. This could certainly affect the recognition of the current word, and consequently, the identification of the onset of the following word. In such situations, this effect probably plays a less important role and the recognition and segmentation rely on other processes. However, this does not exclude the possibility that this local effect (current-word recognition influences the identification of the onset of the following word) is used in HSP most of the time. Such a local effect imposed by the recognition of the current word upon the segmentation of the following word is not directly implemented in ASR where multiple segmentation hypotheses are derived by imposing semantic and syntactic constraints and precede the final recognition of the whole sentence. However, in principle, such an approach could be also implemented in ASR by providing onset markers after recognition of high-confidence words and associating additional probabilities with these segmentation markers based on the confidence measure of the preceding word.

5.6. Coping with Speech Variability

Probably the most difficult problem in ASR is dealing with the great variability found in natural speech and with the effects of various noisy environments. In both cases, current technologies would perform better if more speech training data, covering a larger number of speakers and environmental conditions, were available. However, humans perform much better than ASR in perceiving speech from many speakers and different environments without a prior exposure to the exact type of speech and environmental noise. That leads us to suspect a deficiency in ASR in coping with the large variability in speech and environmental factors.

By way of example, the left plot in Figure 1 shows the distribution, in the acoustic space of the first two MFCCs, of all the frames corresponding to the 61 phones from the training part of the TIMIT database (blue/dark) and to the /aa/ phone (red/grey). Such large spectral distributions for phonemes have been shown to be more the result of phonological context variability than of speaker variability. Speech sounds also display a large temporal (durational) variability. Although a normal speech rate comprises about 10 phonemes per second, speech can be produced and perceived at much lower and higher rates. For example, the comedienne Fran Capo, who is listed in The Guinness Book of World Records as the fastest talking female, has achieved an incredible speaking rate of 603.32 words per minute, which is about 10 words per second. It is more difficult to assess the maximum rate for speech perception than that for speech production, but it is clear that it is much higher than 10 phonemes per second. Another temporal (durational) variability, which undoubtedly plays a significant role in speech perception, is the effect of phonological context on phone duration [41]. The right plot in Figure 1 shows the variability of the average duration across all 61 phones in the training part of TIMIT as a function of right phone identity (written at the bottom). To account for this high degree of variability, there must be intricate mechanisms in the brain. It is likely that the brain deals with the high degrees of spectral and temporal variability of speech by employing specialized neural mechanisms.

In HMM, temporal variability is modelled by transitional probabilities (or by explicit state or phone duration models) among stationary (or non-stationary) phone states and spectral variability is usually accounted for by employing a mixture of a large number of Gaussian densities to represent the composite probability density of the acoustic features of the phones in each state. A mixture of Gaussian is associated with an HMM model of a phone and it provides a spatial representation for this phone model in a P dimensional space of the feature vectors, containing P static and dynamic features such as MFCCs. Each Gaussian mixture can be imagined as a hyper sigmoid (or hyper sphere) in a P dimensional space, but the standard deviations of the features are usually different. The dimension of the space is equal to the dimension of the feature vector and it is usually less than 40. All the phones are represented in this common and homogeneous P dimensional space, each occupying somehow distinct regions, but the corresponding hyper sigmoids (e.g., 5-20 for each phone) are not disjoint and they intersect to a significant extent. Even adding more speech data (and more Gaussian mixtures) does not necessarily lead to a better separation of the hyper sigmoids corresponding to different phone models. In humans, as the information proceeds from the cochlea to higher hierarchical levels in the auditory pathway, an increase in the dimensionality and the heterogeneity of this space takes place (30,000 in the auditory nerve and 500,000 in the medial geniculate body before entering the auditory cortex). Since a neuron can receive information from thousand of other neurons representing various but related information, the neuron can explain how the brain can account for a large variability of sensory information corresponding to a specific category (a brain code). Heterogeneous spaces of representation doubled by parallel neural channels specialized for ranges of variability could explain in part the high performance of HSP in various conditions (speakers and environments). Another explanation could be a highly complex top-down process for integrating not only syntactic-semantic information with the acoustic bottom-up information but also multimodal semantic information from conceptual levels. Hence, in general, in HSP the variability of speech is more likely accounted for by many parallel populations of neurons that map into the same higher level representation of a speech segment (phoneme or phonological sequence), each specialized to cover a specific region of the variability (similar to what a Gaussian mixture does). The distinction is that these speech segments might be represented by heterogeneous multi-dimensional spaces, i.e., using different heterogeneous spaces specialized for particular speech segments and particular ranges of their variability.

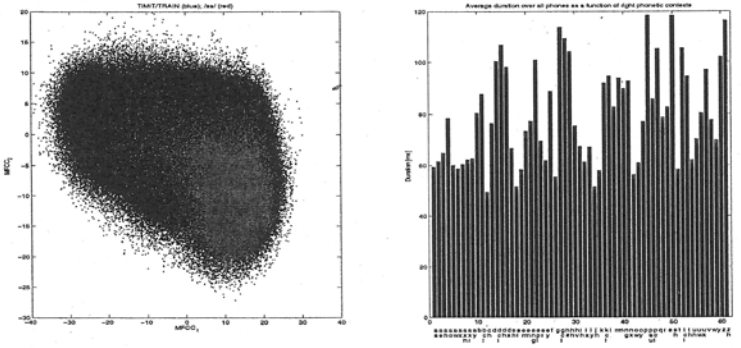

Figure 1. Left: MFCC distribution for all the TIMIT training phones (blue/dark) and the /aa/ phone (red/grey); Right: Average phone duration (ordinate) as a function of the right phone context (abscissa)

6. DISCUSSION

Increased understanding about the brain and language processes has emerged, especially during the last two decades [42]. For speech processing in the brain there is evidence that more complicated architectures and processes are involved than was thought to be the case not so long ago, when it was assumed that speech was only processed in the left hemisphere (in the Broca's area for production and Wernicke's area for perception). There is a better understanding now about how words and concepts emerge and are related in the brain [43], [44]. In particular, this new understanding can have a significant impact upon current and future implementations in ASR.

The detailed time varying processes involved in the perception of spoken words are not yet fully understood, but when this understanding is obtained it could produce a significant step forward in the advancement of ASR technology. It is almost impossible to think about a spoken word without following an underlying imaginary time sequence of the sounds of the word. However, one can think (with a little less difficulty) about a concept without activating to consciousness the corresponding word code. Although recent time course theories for speech production are more accurate in reflecting the real processes in the brain [45], there is still no clear evidence about the exact time course processes involved in lexical access during speech perception.

A related but significant issue is whether there is a fundamental unit in speech perception, and what it is. Indirectly, this would explain whether words are processed holistically or microscopically (by a pre-processing stage which presupposes the recognition and concatenation of sub-word units) in the brain. There might be also a third alternative as suggested in this paper: namely combining the first two alternatives. In such a model, words are perceived both holistically and microscopically (possibly involving various phonological units or features) and the information emerging from various levels is integrated leading to a decision among competing candidates. It is known that humans can accurately perceive isolated words and syllables or some phonemes that can be produced in isolation. That simply means that humans have the ability to perceive various speech 'codes' even without a larger context or without meaning. This does not mean that the brain uses exactly the same processes and channels in all these tasks or that the accuracy of identification is the same in all these cases. However, the alternative suggested here for speech perception is different from that suggested in [23], which supports the idea that there is no fundamental unit in speech perception and the unit is what the attention focuses on in a specific task. The idea suggested in this paper indeed agrees that in the perception of words the objects of attention are the words, whereas in the recognition of nonsense syllables or phones the objects of attention are these units. But, in addition, this paper suggests that the perception of words involves a few parallel processes that all concur in the perception of the word. That is, there are simultaneously involved processes: a holistic process of word perception and a few other sub-word processes such as for the recognition of syllable, diphones, and certain phones. This hypothetical model does not contradict the idea that the perception of words is also influenced by larger lexical segments (phrases, sentences) or semantic information, but the discussion here is limited to the processes at the word level and below. There is some evidence that speech is not represented in the brain during perception as features, but even if it is, this does not contradict the idea suggested in this paper.

The most important argument supporting this multi-level model of word perception proposed here is that all of these units are repeatedly heard during the process of language acquisition, and this must inevitably lead to the creation of architectural patterns of synaptic connectivity in the auditory pathway and the auditory cortex at various hierarchical levels and not only to a single pattern ending with the word 'code'. For example, why should the brain not build a myriad of synaptic connections (since it has so many available during infancy) for the repeatedly heard spectral transitional pattern between /ʃ/_/i/ in words such as 'she', and only build such patterns for words or phonemes? It is known that such synaptic connections and the early beginnings of recognition of words occur during the first year of infancy whereas the acquisition of the meaning of the words only begins during the second year of the child's life [14]. Since the absence of the meaning of words does not preclude the building of such word codes from continuous speech, it is unlikely that the brain does not build similar codes for syllables or even phonemes due to a lack of meaning when they are heard repeatedly during language acquisition.

Another argument espouses the theory that speech is heard and coded in the brain as a concatenation of fundamental units (e.g., phonemes). This theory is contradicted by the fact that the concept of the phoneme is mostly acquired by humans when they learn how to read and write in alphabetical languages. But this does not mean that illiterates are not able to recognize and repeat, for example, an isolated nonsense syllable such as 'bir' or an isolated phoneme such as /f/. If they would only have learned word codes these sub-word units could not be recognized. But, they probably have previously built synaptic patterns that respond to 'i', 'r', 'f' and possibly 'b', and others that respond to 'bi' and 'ir', whether or not these patterns later converged into meaningful words. So, the 'codes' for these units must exist in the human brain together with those of words since infancy and childhood. One can argue against this by saying that humans can repeat things without having previously learned codes for those things. For example, a French-only speaker could repeat an isolated Japanese word without previously hearing that word or knowing its meaning. This is true, but this is only possible because the person previously heard and spoke similar syllables and phonemes in French. When an English-only speaker hears for the first time the Romanian word 'mâine' (meaning 'tomorrow') pronounced as /m ɨ i n e/ (IPA symbols), he is not able to repeat it correctly because the close central vowel /ɨ/ does not have perceptual and production codes in his brain.

Perceptual studies on the spectral transitions between phonemes showed evidence that these regions play a very important role in speech perception [46]. If these transitions are so important why doesn't the brain have individual recognition 'codes' for them, since they are not characteristics of individual phonemes (they do not belong to phonemes) and they only characterize specific combinations between phonemes? Since humans usually retain a lexicon comprising a few tens of thousand of words, what would be the memory economy of not employing a few more thousand codes for these important phonological segments? A recent study showed that the spectral transitions between phonemes reveal quite unexpected underlying trajectories, some of them non-monotonic, which do not have a linear relation to the corresponding monotonic trajectories of the articulators during the production of these segments [47]. Moreover, in a different study, it was found that the spectral variability at the position of maximum spectral transition between phonemes is lower than that at the phoneme centers, and this difference is statistically significant [29]. In addition, there are statistical regularities in the duration of phone as a function of the left or right phone context [41] and these regularities could carry some linguistic information. These findings, among many others, support the idea that speech perception likely uses learned transitional 'codes' in the perception/recognition of words. So, it seems unreasonable to believe that neither the fundamental sounds (phonemes) nor their various phonological combinations have synaptic codes specialized for their recognition. Such a multi-level word perception/recognition process would not contradict the limitations from the information processing capacity of humans since they could be performed in parallel at an unconscious level.

Arguments can also be constructed in support of the holistic process of word recognition. The capacity of the short time memory in humans is much higher than that corresponding to a single word, and it is unlikely that the brain does not also employ holistic channels to perceive the words. These channels could be explained by the redundancy principle involved in perception, and could account for error corrections when the incoming speech is syntactically defective, as in phonemic restoration [48]. In reading, words are processed by saccadic eye movements with fixations of less than 300 ms that correspond to some 4-6 letters, and thus words of this lengths or shorter are processed as a whole [49]. It is likely that a holistic process is also involved in reading, although it might not be the only one. Some researchers suggest that a phonological sequencing is also involved in word reading [50], but this does not exclude the existence of a holistic process and the phonological sequencing could be a parallel process of visual-word perception because of the way most people learned to read.

The multi-level model of word perception could be extended to higher levels of spoken sentence perception by imposing additional processes of real-time segmentation at the word level as suggested in Section 5.5. In addition, a similar real-time segmentation process could take place at the sub-word units level. This model could also be incorporated into adaptive spoken language understanding systems that represent the knowledge at two separate and interactive levels (a lexicon word level and a semantic word level), in addition to the syntactic-grammatical level, and which could also model the acquisition of words and meanings [51]. The multi-level model could also be extended towards lower levels responsible for coping with acoustic variability in which specialized parallel channels represented by heterogeneous spaces could map into the same category code (phone, diphone, syllable, word), as suggested in Section 5.6.

The multi-level model of word perception is well supported by principles of redundancies, which are considered to play an important role in HSP and perception in general. It is also supported by the availability of an immense number of neurons in the auditory sensory system. It is supported by the evidence that the brain has segregated (specialized) areas for the processing of various characteristics of the sensory information (e.g., in vision --- form, color and movement). So what kind of biological principles could contradict such a multi-level model of word perception? The model proposed here is merely a sum of ideas instead of a theory. In no way is it suggested that it is complete or 100% accurate. Preliminary experiments implementing these ideas are underway, but it is premature to expect a definite conclusion from them as yet.

7. CONCLUSION

Speech is certainly perceived by humans by involving more complicated and robust processes than those currently implemented in ASR. It appears that this is a consequence of the enormous number of 'information' carriers (~100 billion neurons) and of the even higher number of interconnections (~100 trillion synapses) present in the brain. However the number of these microscopic elements is not the only key difference: the brain relies on many parallel and redundant channels that converge various information representations into cognitive categories, and these channels are specialized for specific attributes using many hierarchical levels of processing. By comparison, the existing models of ASR only implement a relatively small number of hierarchical levels where the information is processed in a parsimonious manner, due to computational constraints. In ASR the acoustic and language models appear to be more inspired from phonetics and linguistics than from neurobiology of speech perception. Although current microprocessors are driven by clocks running at a few GHz, this speed cannot account for the enormous number of elementary processes and interconnections that take place in the brain, mostly in parallel. Even if future advances in computing could bridge this computational gap, this would still not be enough to reproduce or match human performance in a task such as speech perception. It is anticipated that significant advances in the understanding of how such brain processes work together to enable HSP would provide sufficient insight into bridging the performance gap between HSP and ASR performance, even in the most adverse and noisy environments.

8. ACKNOWLEDGEMENTS

This work was supported in part by the U.S.A. National Science Foundation under the Knowledge and Distributed Intelligence project (grant NSF IIS-98-72995) and the Automatic Speech Attribute Transcription project (grant NSF IIS-04-27431).

REFERENCES

1. LIPPMANN, R. P., Speech recognition by machines and humans. *Speech Communication*, **22**, pp.1–15, 1997.
2. PIERCE, J. R., Whither Speech Recognition? *The Journal of the Acoustical Society of America*, **46** (4) part 2, pp. 1049–1051, 1969.
3. HARPER, M., National Science Foundation (NSF) Symposium on Next Generation Automatic Speech Recognition, http://www.ece.gatech.edu~chl/ngasr03/, Atlanta, GA, Oct. 7–8, 2003.
4. WAYNE, C., Defence Advanced Research Project Agency (DARPA) - Effective, Affordable, Reusable Speech-to-Text (EARS) Program. http://www.darpa.mil/ipto/programs/ears/.
5. WORDEN, F. G., Hearing and the neural detection of acoustic patterns. *Behavioral Science,* **16**, pp. 20–30, 1971.
6. SHEPHERD, G. M., *Synaptic Organization of the Brain,* Oxford University Press, 1998.
7. LYON, R. F., A Computational Model of Filtering, Detection, and Compression in the Cochlea, in Proceedings of IEEE-ICASSP-82, pp. 1282–1285, 1982.
8. SENEFF, S., A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics,* **16** (1), pp. 55–76, 1988.
9. HERMANSKY, H., Perceptual Linear Predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, **87** (4), pp. 1738–1752, 1990.
10. GROSSBERG, S., How Does a Brain Build a Cognitive Code? *Psychological Review,* **87**(1), pp. 1–51, 1980.
11. SPERBER, D., and WILSON, D., The Mapping between the Mental and the Public Lexicon. *Thought and Language.* Cambridge University Press. pp. 184–200, 1998.
12. BOLTE, J. and COENEN, E., Is Phonological Information Mapped onto Semantic Information in a One-to-One manner? *Brain and Language.* **81**, pp. 384–397, 2002.
13. DAMASIO, A. R. and DAMASIO, H., Brain and Language. *The Scientific American Book of the Brain.* The Lyons Press. pp. 29–41, 1999.
14. JUSCZYK, P. W., How infants begin to extract words from speech. *Trends in Cognitive Science.* **3**(9), pp. 323–328, 1999.
15. MCGURK, H., and MACDONALD, J., Hearing lips and seeing voices. *Nature.* **264**, pp. 746–748, 1976.
16. MASSARO, D. W., and ODEN, G. C., Independence of Lexical Context and Phonological Information in Speech Perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition* . **21**(4), pp. 1053–1064, 1995.
17. BOFF, K. R., KAUFMAN, L., and THOMAS, J. P., (Eds.), *Handbook of perception and human performance.* Wiley, New York. 1986.
18. RAMSEY, N. F., JANSMA, J. M., JAGER, G., VAN RAALTEN, T. AND KAHN, R. S., Neurophysiological factors in human information processing capacity. *Brain.* **127**, pp. 1–9, 2004.
19. MILLER, G. A., *The Science of Words*, New York, Scientific American Library, 1991.
20. FLANAGAN, J. L., *Speech Analysis Synthesis and Perception*, Springer-Verlag, Berlin, Heidelberg, New York, 1972.
21. GREENBERG, S., CARVEY, H., HITCHCOCK, L, and CHANG, S., Temporal properties of spontaneous speech --- a syllable-centric perspective. *Journal of Phonetics.* **31**, pp. 465–485, 2003.
22. RABINER, L., JUANG, B.-J., *Fundamentals of Speech Recognition*, Englewood Cliffs, New Jersey, Prentice Hall, 1993.

23. GOLDINGER, S., and AZUMA, T., Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics.* **31**, pp. 305–320, 2003.
24. PLOMP, R., *The Intelligent Ear: On the Nature of Sound Perception,* Lawrence Erlbaum Associates, Mahwah New Jersey, London, 2002
25. WARREN, R. M., *Auditory Perception: A New Analysis and Synthesis*, Cambridge University Press, Cambridge, 1999.
26. STEVENS, K, N., and BLUMSTEIN, S. E., Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, **64** (5), pp. 1358–1368, 1978.
27. BLUMSTEIN, S. E., and STEVENS, K, N., Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, **66** (4), pp. 1001–1017, 1979.
28. KEWLEY-PORT, D., Time-varying features are correlates of place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, **73**, pp. 322–335, 1983.
29. DUSAN, S., Spectral variability at the transition between successive phonemes. *The Journal of the Acoustical Society of America*, **115** (5), Pt. 2, 1pSC27, pp. 2428, 2004.
30. SAVIN, H. B., and BEVER, T. G., The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior.* **9**, pp. 295–302, 1970.
31. LIBERMAN, A. M., and MATTINGLY, I. G., The motor theory of speech perception revisited. *Cognition.* **21**, pp. 1–36, 1985.
32. DIEHL, R. L., and KLUENDER, K. R., On the objects of speech perception. *Ecological Psychology.* **1**, pp. 121–144, 1989.
33. NEAREY, T., Speech perception as pattern recognition. *The Journal of the Acoustical Society of America*, **101**(6), pp. 3241–3254, 1997.
34. MCCLELLAND, J. L., and ELMAN, J. L., The TRACE model of speech perception. *Cognitive Psychology.* **18**, pp. 1–86,1986.
35. MASSARO, D. W., Testing between the TRACE Model and the Fuzzy Logical Model of Speech Perception. *Cognitive Psychology.* **21**, pp. 398–421, 1989.
36. JUSCZYK, P. W., A review of speech perception research. *Handbook of perception and human performance.* New York: Wiley, pp. 27–57, 1986.
37. GREENBERG, S., Speaking in shorthand --- A syllable-centric perspective for understanding pronunciation variation. *Speech Communication.* **29**, pp. 159–176, 1999.
38. ROUILLER, E., Functional Organization of the Auditory Pathways. *The Central Auditory System.* New York, Oxford, Oxford University Press, pp. 3–96, 1997.
39. WAIBEL, A., HANAZAWA, T., HINTON, G, et al., Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing* . **37**, pp. 328-339, 1987.
40. ZEKI, S. The Visual Image in Mind and Brain. *The Scientific American Book of the Brain.* The Lyons Press. pp. 17–28, 1999.
41. DUSAN, S., Effects of phonetic contexts on the duration of phonetic segments in fluent read speech. International Conference on Spoken Language Processing, ICSLP'04 Proceedings, S. Korea, 2004.
42. DAMASIO, A. R., Brain and language: what a difference a decade makes. *Current Opinion in Neurology.* **10**, pp. 177–178, 1997.
43. GROSSBERG, S., Resonant neural dynamics of speech perception. *Journal of Phonetics.* **31**, pp. 423–445, 2003.
44. DAMASIO, H., TRANEL, D., GRABOWSKI, T., ADOLPHS, R., and DAMASIO, A., Neural systems behind word and concept retrieval. *Cognition* . **92**, pp. 179–229, 2004.
45. LEVELT, W. J. M., ROELOFS, A., and MEYER, A. S., A theory of lexical access in speech production. *Behavioral and Brain Sciences,* **22**, pp. 1–75, 1999.
46. FURUI, S., On the role of spectral transition for speech perception. *The Journal of the Acoustical Society of America*, **80** (4), pp. 1016–1025, 1986.
47. DUSAN, S., Non-monotonic spectral transitions between successive phonemes. *The Journal of the Acoustical Society of America*, **116** (4), Pt. 2, 1pSC2, pp. 2479, 2004.
48. WARREN, R. M., Perceptual restoration of missing speech sounds. *Science.* **167**, pp. 392–393, 1970.
49. JUST, M. A., and CARPENTER, P. A., *The Psychology of Reading and Language Comprehension.* Allyn & Bacon, Boston, 1987.
50. XU, B., GRAFMAN, J., GAILLARD, W. D., SPANAKI, M., ISHII, K. BALSAMO., L., MAKALE, M., and THEODORE, W. H., Neuroimaging reveals automatic speech coding during perception of written word meaning. *NeuroImage,* **17**, pp. 859–870, 2002.
51. DUSAN, S., and FLANAGAN, J., Adaptive Dialog Based upon Multimodal Language Acquisition, The Fourth IEEE International Conference on Multimodal Interfaces, Pittsburgh, PA, USA, pp. 135–140, 2002.