Aspects on speech perception as the model for ASR

GSLT speech technology course | fall 2004

Lisa Gustavsson
Department of Linguistics, Phonetics
Stockholm University, S-106 91 Stockholm, Sweden

Abstract

Over the last decades it has become increasingly popular to adopt inspiration from knowledge on human language skills in applications within speech technology. It has proven to be successful in some aspects, such as signal processing in automatic speech recognition[1] (ASR) but the overall performance is still far behind that of humans. Perhaps problems within speech technology can be solved with improved engineering but a closer collaboration between technology and cognitive science might lead to insight on more fundamental principles involved in language processing that could turn out useful in speech technology.

1. Introduction

This paper presents a theoretic discussion of problems concerning automatic speech recognition and how knowledge about human speech perception is used as inspiration for finding solutions to these problems. Automatic speech recognition is used in many applications today and it works well provided the acoustic setting is adequate and it does not encounter new or unexpected input. Humans however, have an astonishing skill for coping with such situations; if we do fail in our speech perception task we have the ability to recover by creating hypotheses of what is being uttered and if we encounter new situations we have the capacity to learn. This remarkable ability to perceive and understand speech has given rise to inspiration when designing ASR systems. Knowledge about human speech perception is in some aspect incorporated in every state of the art system (for example in speech processing products from CISCO, AT&T, Apple and Microsoft to mention a few) but some pieces are still missing since the improvement is marginal. Simulating humans' ability to perceive speech is a brilliant idea, but problems seem to arise when this ability is taken out of its context and implemented in an ASR system. It becomes obvious that there is a lot more to speech perception than signal processing. Speech perception in humans[2] is a skill that takes years to achieve and the process is intertwined in the cognitive web and highly dependent on sensory input other than the auditory. Is it really necessary to simulate the entire realm of human intelligence in order to make systems that perceive speech satisfactorily or should speech technology be primarily concerned with finding the best shortcuts and settle with systems that are almost good enough?

First a brief summary of the basic architecture of current automatic speech recognition systems (section 2) and an introduction to various problems concerning ASR (section 3) that will serve as a framework within which speech processing will be considered throughout the paper. This is followed by an overview of how automatic speech processing could be viewed in light of human speech perception (section 4). This matter is then discussed in more detail in relation to theories and ideas on speech technology and theories on learning and development in general (section 5). Finally, some new trends within cognitive science and robotics that concerns speech perception are presented (section 6). The concluding remarks (section 7) summarise the paper and provide some speculations on the future of ASR.

2. Brief ASR overview

There are various types of ASR systems, ranging from one speaker recognisers to multiple speaker recognisers and from specific domain systems (such as digit recogniser or timetables) to large vocabulary systems. There used to be a distinction between continuous speech and discrete speech recognition, but as a consequence of a more demanding market and more powerful computational recourses most of today's systems are designed to recognise continuous speech. The task of any ASR system is translating continuous acoustic signals to linguistically intelligible representations of speech. The techniques used in current ASR systems for conducting this task can be described in three basic steps. First, feature extraction, methods for analysing and describing acoustic signals as discrete segments labelled according to their acoustic properties. Second, pattern recognition, where acoustic models are used to calculate transition probabilities between acoustic segments according to a pronunciation lexicon. Third, linguistic scoring, statistical grammar models according to vocabulary and syntax rules are used for calculating probabilities of word sequences.

[1] Throughout the paper I will discuss aspects of Automatic Speech Recognition in general and use the term ASR, but include many aspects of ASU (automatic speech understanding) in this label.
[2] Indeed other mammals too are excellent in perceiving speech, but in this paper I will focus on the human ability to perceive in order to understand a linguistic message.

2.1 Signal analysis

Feature extraction methods are very sophisticated in the sense that they make use of all acoustic information available in the waveform. Steady regions in the speech signal are categorized according to their acoustic properties (such as formant locations, nasality, frication, voiced and unvoiced). This is usually done by using either Fast Fourier Transform analyses (FFT) that break down the complex waveform into its discrete frequency/amplitude components such as harmonics in a speech signal, or Linear Predictive Coding (LPC) that is a method for discovering regularities in any time varying data, such as formants in a speech signal. An LPC analysis is *predictive* in the way that it tries to predict the upcoming value based on the hypothesis that any sample in the waveform is a direct function of the preceding sample. This reduces the search space considerably but the output is a rather crude estimation of resonance components without information about the individual harmonics (as in an FFT analysis). But, since harmonics are only interesting for identifying the fundamental frequency but not necessarily for discovering formants, LPC is a good solution for defining the characteristics of different speech sounds. For a more comprehensive description of techniques for speech processing, see Kent and Read (1992), Johnson (2003) or Stevens (1998).

2.2 Pattern recognition

No region in the speech signal is truly stable, but rather a reflection of the dynamic vocal tract filter configurations. It is therefore necessary to take the acoustic context into account when labeling the segments. The acoustic models used in pattern recognition are usually based on this knowledge about speech production and reflect the behavior of the articulators. According to these acoustic/phonetic rules (that are either learned from training or preprogrammed) Hidden Markov models (HMMs) are used to calculate transition probabilities from one acoustic segment to another to convert the segments into phones (or some other appropriate symbol such as syllable or phoneme) selected from a pronunciation lexicon (Rabiner, 1989). The original input speech signal is now represented by a sequence of phones from the pronunciation dictionary. Each phone comes with a context dependent HM model for calculating transition probabilities between different phones. The HMM outputs are strings of phones that constitute possible sub-word or word candidates with a certain probability.

2.3 Language modeling

This final step is built upon linguistic knowledge. The sequence of word candidates is parsed according to language specific rules. Such rules can either be derived statistically from extensive training with similar input data or from grammar and syntax specifications of the language in question. These rules constitute a language model that, given the adjoining words, calculates the probability of a certain word occurring in that particular context and hence determine the most probable word sequence.

2.4 Optimisation

Today's systems do have excellent signal processing modules and built in knowledge sources of the language in question, but this is not sufficient for handling the enormous acoustic variety in different input signals. There need to be a dynamic component for learning processes that enables the system to adjust to new unexpected events. At least when developing an ASR system, training is essential not only for deriving implicit linguistic rules that correspond to the nature of input the system is designed for, but more so for gaining insight about the acoustic properties of a variety of speech and non-speech signals under different conditions. This strive for optimisation is facilitated by the methods used in today's system, such as Hidden Markov that simplifies learning processes by continuously re-evaluating weights during training. However, while stringency punishes tolerance, the price paid for flexibility is errors due to generalisations. Built in acoustic speech production models and language models that reduce the search space and restrict outputs together with dynamic programming and statistical models that offers flexibility to the system constitute the compromise many of today's ASR systems are built upon.

3. Performance

The problem facing anyone who is involved in developing an ASR system is how to deal with the fact that one linguistic message can come in so many different shapes. Different pronunciations between speakers and within the same speaker, but in different situations, are endless in their varieties. The goal of speech technology is human-like performance, hence comparisons between machines and human subjects as measurements for progress are highly motivated within automatic speech recognition. Lippmann (1997) investigated a number of such evaluations and found that error rates were consistently significantly higher in ASR than in humans. Although the speech perception performance of humans is the ambition, humans do have exactly the same problem as ASR systems but not to the same extent. Noisy background, poor transmission conditions (such as distorted frequency over a telephone line), distance, different speaking styles, accents and out of context utterances are some examples of where the human listener, just like any ASR system, sometimes fails. Still, if pure engineering methods of signal processing are not sufficient for performing speech signal analyses,

inspiration has to come from knowledge about human speech perception. Humans have an amazing ability for error-recovery and the spoken language works very well for communication. Therefore this is what we are aiming at in an ASR system. This requires that the system is flexible enough to readjust phonological representations, able to adapt and learn new lexical, syntactic, semantic and pragmatic information as a human can.

4. Speech perception as inspiration

In the brief ASR overview above the speech recognition process seems very technical, but inspiration from human speech perception pervades many of the methods used in today's systems. Knowledge concerning anything from how the ear transmits the acoustic signal to semantic interpretation is implemented to recreate what hopefully is the intended phonological underlying representation of the input signal. The reasoning is this: Since the characteristics of language originate from constraints due to the human speech production- and speech perception apparatus (Lindblom, 2000), developing a system with the same constraints might facilitate the process of uncovering the raw signal to get to the fundamental linguistic message.

4.1 Signal analysis in human periphery hearing

In many aspects some general parallels can be drawn between the methods used in acoustic signal analysis and human speech perception. LPC for example, was developed for efficient tracing of regularities in the signal, such as broad spectral peaks. In the same way humans (and many other mammals too for that matter) are very sensitive to spectral regularities in the signal such as formant locations when distinguishing between different speech sounds. Statistical methods such as HMM for calculating transition probabilities also work pretty much the same way as a skilled language user that exactly knows what a particular speech sound should sound like in relation to adjoining sounds. Another example is the well known logarithmic Decibel conversion of sound pressure. Sound pressure is usually expressed in Pascal which is a linear scale, but when it comes to speech perception Decibel is a more suitable scale since it represents the nonlinear relation between sound pressure and the perceived loudness. There are other similar scales for perception of loudness which are also used in applications of speech perception, such as the Sone-scale (Stevens, 1957) that is purely based on empirical data from a number of subjects.

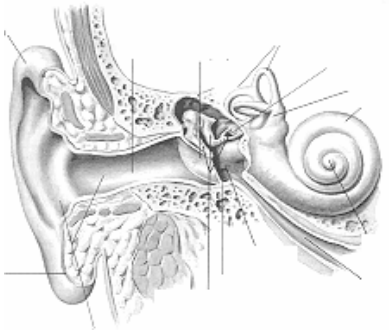

Figure. 1 (above) Human ear. From Sonesson & Sonesson, 1993.

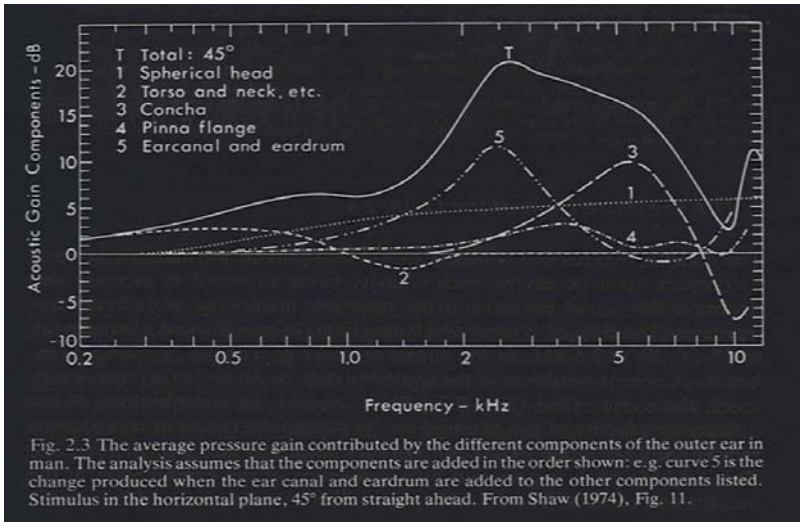

Fig. 2.3 The average pressure gain contributed by the different components of the outer ear in man. The analysis assumes that the components are added in the order shown: e.g. curve 5 is the change produced when the ear canal and eardrum are added to the other components listed. Stimulus in the horizontal plane, 45° from straight ahead. From Shaw (1974), Fig. 11.

Figure. 2 (right) Average sound pressure gain in the human ear. From Pickles, 1988.

Indeed, also the initial acoustic analysis of the speech signal can be viewed as the function of the peripheral auditory system (see figure 1). The acoustic signal gets modulated when entering the external ear canal which has the same filter functions as an acoustic tube closed at one end (the eardrum) and open at the other. Due to these resonance properties of the ear canal, but also resonance properties of the head, the torso and the ear conch, some frequencies in the signal will be amplified, for example frequencies around 2-3 kHz gain about 15-20 dB (see figure 2 and Pickles, 1988). The pressure gain curves in figure 2 indicate how different frequencies are weighted to achieve the perceived subjective loudness. This relation between frequency and loudness can be expressed in terms of *the Fletcher-Munson equal-loudness curves* where hearing-thresholds for all frequencies in the range of human hearing capabilities (approximately 0.025-20 kHz) are estimated. This phenomenon is something that early was considered in ASR (Itahashi and Yokoyama, 1976) and Fletcher-Munson equal-loudness curves are today incorporated in some of the most common methods for speech analysis. One of these is Hermansky's (1990) Perceptual Linear Prediction (PLP) where this asymmetrical sensitivity at deferent frequencies is imitated by subtracting these hearing thresholds from the input signal. The result is a signal that appears like a signal that has passed through the human outer-ear. Since characteristics of speech sounds are mainly defined

by frequencies around 1-4 kHz, this is a relevant perceptually-based method for capturing central information in the speech signal.

The signal undergoes some changes when travelling through the chain of bones in the middle-ear as well, but these transformations have little impact on the actual speech perception process. In the inner-ear on the other hand is where the spectral analysis of sound takes place. The basilar membrane (see figure 3) in the shell-shaped cochlea (see figure 1), is the headquarter of frequency analysis. In some respects the function of the basilar membrane could be compared to that of a Fourier Transform (Yates, 1993). One end of the basilar membrane is stiff and responds to high frequency oscillations, further down the membrane gets softer and responds to low frequency oscillations. Thousands of tiny hair cells along the membrane (in the organ of Corti, see figure 4) transform these oscillations to impulses to the auditory nerve. The signal has been analysed in its frequency components and the information is sent to the brain for further investigation.

However, psychoacoustic studies have shown that this frequency response in the basilar membrane is more or less logarithmic and dramatically influences our subjective interpretation of the speech signal (Gelfand, 1998). This frequency resolution can be described in terms of critical bands where each band corresponds to a certain continuously increasing frequency width. In contrast to the physically grounded and linear signal analysis performed with FFT or LPC, the mapping of frequency in the basilar membrane is not a direct match with the acoustic signal. The critical bandwidths can be said to function as filters, which to a certain extent enhance certain sounds (for example, our subjective perception of loudness changes when the distance in frequency of two signals exceeds such a bandwidth whereas the perceived loudness is even within one critical bandwidth). While the critical bandwidth expressed in Hz is quite narrow in the lower frequencies (the soft apical end of the basilar membrane, by the helicotrema) and small changes in frequency are detected, further up in frequency the bandwidth gets broader and it takes bigger changes in frequency in order to perceive them.

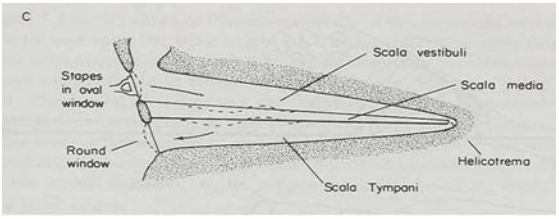

Fig. 3 (above) The path of vibrations in the basilar-membrane are shown in a schematic view of the cochlear duct unrolled. From Pickles, 1988.

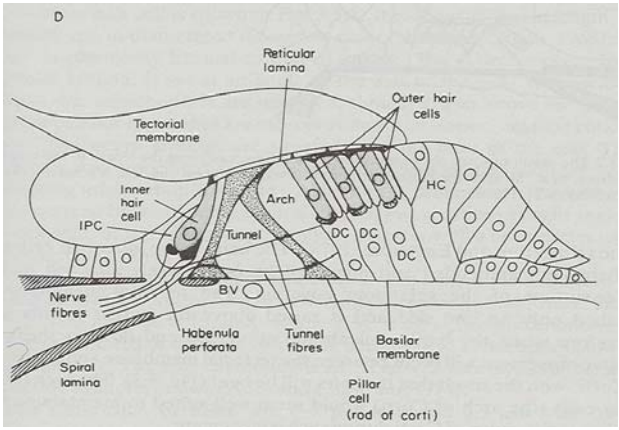

Fig. 4 (right) A cross-section of the organ of Corti. From Pickles, 1988.

The discovery of this nonlinear relation between perceived frequency and Hz gave rise to a couple of perceptually based frequency scales called the Bark scale and the Mel scale. One Bark is defined as one critical bandwidth and corresponds roughly to a step of constant length along the basilar membrane, i.e. one Bark corresponds to an approximately constant number of hair cells in the basilar membrane. The Mel scale is very similar to the Bark scale in that it is based on the subjective perception of pitch and transforms the signal to correspond to this nonlinear human perception of frequency. This kind of knowledge about frequency resolution in human peripheral hearing is something that is used in many of today's techniques for analysing speech signals. The most popular one is the Mel Frequency Cepstrum Coefficients (MFCC) method, where a Mel-scale frequency warp is used for converting the input-signal to what is thought to be its auditory representation. In the same way, in the PLP-analysis (Hermansky, 1990) mentioned earlier, the Bark-scale (the critical band spectral resolution) in used for filtering the signal. The techniques are similar except for the exact shape of the frequency-perception curves.

4.2 Linguistic knowledge

Leaving the signal processing step of speech recognition and entering the linguistic modelling more extensive descriptions of language rules are implemented in the system. Any parallels to the human speech perception here must be drawn to a cognitive level rather than auditory. Extensive work has been done also in this area to implement modules of cognitive linguistic capabilities, such as grammar, semantics, pragmatics and syntax in ASR systems. Taken together, these components added to the thorough initial acoustic analyses and categorisations, fused by final evaluations of suprasegmental aspects such as intonation, speaking rate and speaking style, seem like the perfect imitation of the human language skill in such a way that the system uses a number of knowledge sources and relevant methods in order to derive the linguistic message embedded in an acoustic signal.

Apparently this cannot be true, as mentioned in section 3, humans are still superior to machines in all aspects of speech recognition. The majority of scientists involved in speech technology agree that even the performance of state of the art systems is far below satisfaction and devote their research, now more than ever, to uncover the mystery of speech perception. In the following chapters some of their ideas will be discussed.

5. Is speech perception a proper model for ASR?

There is of course the possibility to construct an ASR system on a purely engineering basis, but inspiration usually comes from the human language skill. The idea of mimicking the human speech perception process is obviously not very farfetched since it is only humans that make use of human language and it is only human language that is the application for ASR. Since acoustic spectral representations of speech tells us something about what the signal sounds like when produced, auditory spectral representations of speech is probably more suitable if we are interested in what the signal sounds like when perceived (cf. Carlson & Granström, 1982). This is a popular view today and many of the systems are brilliant copies of every aspect we know about human speech perception. But there are still problems within ASR.

5.1 Improving the technology

To achieve this human-like speech recognition skill speech technology has gone from signal processing methods to artificial intelligence. Spectrogram was a breakthrough and phonetic insights together with modules of higher-level linguistic competence took ASR to where it is today. Nevertheless, since this does not work to satisfaction we have to track down the weak link in the ASR method chain. According to Ney (2003), the strong standing component over the years in ASR systems is the usage of statistical methods (such as HMM for acoustic matching) but he argues that it is necessary to incorporate proper acoustic/phonetic knowledge in these models, this is something we are lacking today. Similarly Lippmann (1997) is of the same opinion that the stage of low-level acoustic/phonetic analyses is where ASR systems still lack competence and this is where further research is needed. It is true that humans have an amazing ability for detailed acoustic analysis, but so do also the ASR systems today. One guess would be that it is not necessarily the acoustic analysis that has to be improved, but rather the ability to use higher level information for making the correct acoustic analysis. Humans are constantly exposed to a variety in acoustic signals that still get the same interpretation (such as same words but different speakers) or the same acoustic signals that get interpreted in different ways depending on the context (compare for example the phrases *lesson five* and *less than five,* Lindblom, pers. comm.). The weak link in this case would be the kind of knowledge a system uses in order to decide on the right phonological representation of the signal. Batliner *et al*. (2001) believe that this is the case. Today's systems use an intermediate phonological level (pattern recognition) of description between the abstract function (linguistic representation, such as words in the lexicon) and the concrete phonetic form (phonetic features in the speech signal). They argue that better results could be achieved if the systems were functioning successfully without such an intermediate level, using instead a direct link between the syntactic/semantic function and the phonetic form. Such systems would be able to store new knowledge on their own, in a similar way humans do. The intermediate translation to phonologic representations is the main problem according to Greenberg (1997) as well. He suggests that instead of focusing on finding a few linguistically relevant units (such as phonemes) in the speech signal, the system should be able to derive linguistic information from many different sources of information. One single source will not be sufficient for deciding on robust representations of linguistic information, but an analysis of many cross correlated parameters in the speech signal, ranging from acoustic features of a few milliseconds to suprasegmental features stretching over an entire utterance, would be a flexible yet robust method similar to the kind of strategy human listeners use for interpreting the speech signal.

5.2 Non-auditory speech processing in humans

If we were to forget about the signal per se for a moment and examine other aspects of the speech perception task it is immediately obvious that problems concerning decoding a speech signal extend far beyond auditory processing. One main difference between an ASR system and a human language user lays in their adaptive ability – while the ASR system from the start is a fully equipped speech recognizer, it takes years for a human to become skilled in speech perception. On the other hand, humans have throughout the lifespan access to external references (and internal) to their linguistic knowledge that enhance the ability to recover from errors and continuously learn new speech processing tasks. Humans use multiple sources of information at all times for decoding the speech signal, hence a perfect acoustic setting, invariance in the speech signal or a flawless speech perception mechanism is not crucial in humans, whereas in ASR it is. Humans have an awareness of the situation that might be a topic of conversation and have expectations on what is being uttered and interpretations of speech signals are based on such expectations. It is probably not a coincidence that domain-specific ASR systems have the best performance rate. The situation in domain-specific recognition is very much like that of humans, the system knows from the beginning what linguistic message to expect which makes it easy to spot the keywords almost regardless of their acoustic quality. Humans get this information continuously during the lifespan from sensory and cognitive sources while an ASR system has to be pre-programmed with this kind of information.

Another aspect is the setting during the speech act; humans often have access to other linguistically relevant information, such as visual input of the face of the person who is speaking, usually humans don't have to rely solely on the speech signal to solve the signal processing task. Scientists involved in designing dialogue systems are well aware of the fact that such information is crucial in many situations for successful speech recognition and it becomes more and more common to add a face correlated with speech synthesis systems to enhance intelligibility. The other way around, to provide the system with complementary visual input, is not really an option for ASR systems since it would involve simulation of yet another enormous dimension of human perception, that of visual interpretation. The notion of multiple dimensions however, is relevant when constructing human-like ASR systems and needs to be considered in the long perspective. In today's ASR systems all potential linguistic information is squeezed out of the acoustic dimension while humans use information from a multiple dimensions. As discussed earlier, many of the up to date systems do include representations of different knowledge sources to a certain extent, but there are problems with the linking between the acoustic signal and higher-level information.

The conclusions to be drawn from all this seem to be that signal processing per se is probably not the critical issue, with the engineering skills and insights on human speech processing behind today's system, signal processing shouldn't be a problem. The knowledge sources, representing the cognitive high-level processes of human speech perception, incorporated in many of the systems are also very extensive and well motivated. Once again it comes down to the collaboration between signal processing and higher-level interpretation. This however, is a scientific domain which is still in its dawning and efforts from a variety of research areas are necessary to get a better understanding of what could be summed up as *intelligence*.

5.3 Intelligent systems

Minsky (1985) defines intelligence as looking for causal explanations when failing with a task, and when finding them, adding them to the cognitive network of belief and understanding. Intelligent learning develops with experience, and since a task-specific programmed system does not have any experience it cannot learn. Minsky argues for the necessity of childhood in any intelligent system and questions why we make systems do adult tasks before we make them do childish things. This would perhaps be relevant for ASR systems if our demands of them are to perform speech recognition in a human-like fashion. The idea however, of raising an ASR system from a naïve data processor to a fully equipped language user in order to make it recognise speech satisfactorily seems a bit awkward. Nevertheless, the design in today's systems where modules of different knowledge sources are implemented to solve speech processing tasks rests upon traditional viewpoints in the spirit of nativistic theories. The reasoning is that unless humans have some form of innate linguistic knowledge, such as a genetically programmed module enabling language learning and given the variation in speech, it would be impossible for the infant to find any linguistic structure in the varying information flow of the ambient language (Chomsky, 1988, Pinker, 1994).

The reality of such modules of knowledge is not that evident according to Elman (1993) who simulates infants' behaviour in computational network models. He suggests that one does not necessarily has to view the information flow as a problem children must struggle with when learning a language, but instead one should view variation as an asset. In his view, infants have an immature memory and can therefore only process simple sentences and structures (Elman, 1999). This reasoning implies that an infant's internal representations of language are much rougher than adults' in the sense that they are not focused on categories that will be linguistically adequate later on. Elman's suggestion is that improvement in memory capacities and the acquisition of basic phonetic representations allows the infant to process more complex sentences. The input data is also of great importance when the initial capacity of a system is limited but matures with time. Elmans network models[3] imply that lack of variation in the information may cause the system to make wrong generalisations and too much variation in the information may slow down the learning process until enough data is gathered to make generalisations. These findings are in line with results from experiments on sound-meaning connections in infants (Koponen et al., 2003). Another simulation of language learning has been done by Rumelheart and McClelland (1994). Their model, Parallel Distributed Processing (PDP), learns syntactic rules solely on the basis of statistical regularities in the input signal. They chose to investigate English verbs since it is usually claimed that children cannot learn their inflections if they have not acquired the rules of regular and irregular verbs. Their PDP model takes no notice of this, instead it stores all kinds of verbs in a memory that reinforces similar patterns. After a certain amount of input verbs the system finds regularities in the structure of the verbs and generalises them to novel verbs. There are no rules that determine the correct inflection but statistical relationships among the base forms of the verbs. In their view children do not have to find out the rules that could describe a language in order to master it. With this learning model they want to show that rules are all right for describing a phenomenon but do not necessarily have anything to do with the underlying processes of that phenomenon.

[3] For exercises in network modelling, see Plunkett and Elman, 1997.

Humans develop the skill for language processing in many different stages and it appears to be a long and winding road to go from rookie to a flexible, high-performance language user. In ASR systems on the other hand, only that final step with knowledge of an adult language user is implemented and perhaps this is where speech technology fails. This is the opinion of the new branch of computer scientists and engineers that are involved in developing humanoids (robots used for interaction with humans in human environments). We are looking into this in next section while asking the question: Is cognitive development something speech technologists should be concerned with?

6. The merging of science and technology

A theory of mind (specific cognitive ability to understand oneself and others as intentional agents) is believed to be unique for humans, just as language and the two of them seem to go hand in hand during the early developmental stages in humans. Therefore, in quest for insight on the human language, the infants' perspective is something that is taken into consideration and modelled in the field of developmental, intelligent, humanoid robots. In contrast to traditional AI systems, which from the beginning implement different knowledge modules, humanoids build their own knowledge during their life span and the knowledge sources get interconnected in a functional way. A task-specific system (such as ASR) requires human programmers to fully understand the domains of the tasks and to be able to predict them. The developmental approach on the other hand, is motivated by human cognitive and behavioural development from infancy to adulthood. With humanoids, scientists aim to provide a system with broad and unified developmental skills instead of separate knowledge modules. The learning method in a developmental system is simulated by systematic self-organisation processes which enable the system to develop its own cognitive and behavioural skills through direct interactions with its ecological context (the physical world).

One example of how this approach may be useful for gaining insight on language learning is a perceptually grounded robot developed by Domineys and Bouchers (2004). The system is based on a minimum of pre-wired functionalities and learns sound-meaning representations and grammatical constructions by interacting (visually and auditory) with its environment. The language learning behaviour observed in the robot seems compatible with the early stages of human language acquisition. Another example in the same line is the robot SAIL (Self-organizing, Autonomous, Incremental Learner) developed by Weng and Zhang (2002). Also here the baseline is to program only biological conditions such as memory capacities, vision and a highly sophisticated hearing system (Weng and Zhang, 2001). Their overall objective with SAIL is to study how biologically inspired systems potentially can develop intelligence through interaction with their environment (Chen and Weng, 2004). The collective aim within this branch of science is to get an understanding of the human brain through embodied modelling and in the long run be able to construct intelligent systems that are able to function in human environments.

7. Summary

Is it really necessary to understand every aspect of human intelligence in order to make systems that perceive speech satisfactorily? Or should the role of speech technology primarily be about finding the best shortcuts to human-like performance with more fine tuned adjustments of the existing systems? As discussed in the beginning of the paper, ASR systems today work quite well as long as the acoustic setting is adequate and they don't encounter new or unexpected input. The first step of signal processing as well as the last step of linguistic scoring don't seem to be problematic, rather how to intertwine these processes in a functional way seems to be the concern to blame the poor performance (compared to humans). How to solve this problem is the main topic of research within ASR. Whether the pursuit of a flawless speech recogniser takes the long route, through the area of cognitive development, or continues on the track of refining the existing systems by developing the techniques, developmental science and technology are bound to have a closer relationship in the future (Greenberg, 2001). In fact, without speech technology, understanding of human speech processing would not have reached this far, and without empirical knowledge of human speech processing and insight on the power of speech communication, speech technology might not even be justified.

8. References

Batliner, A., Möbius, B., Möhler, G., Schweitzer, A. and Nöth, E. (2001). *Prosodic models, automatic speech understanding and speech synthesis: towards the common ground.* Eurospeech 2001 – Scandinavia.

Carlson, R. and Granström, B. (1982) Towards an auditory spectrograph. In *The Representation of Speech in the Peripheral Auditory System.* Carlson, R. and Granström, B. (Eds.) pp. 109-114. Amsterdam: Elsevier Biomedical Press

Chen, Y. and Weng, J. (2004) *Developmental Learning: A Case Study in Understanding "Object Permanence".* Proc. of Fourth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems. Genoa, Italy.

Chomsky, N. (1988) *Language and Problems of Knowledge* MIT Press, Cambridge, MA

Dominey, P. and Boucher, J. D. (2004) *Developmental stages of perception and language acquisition in a physically grounded robot.* Proc. of Fourth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems. Genoa, Italy.

Elman, J. L. (1999) The Emergence of Language: A Conspiracy Theory. MacWhinney, B. (Red.) *The emergence of language*. Lawrence Erlbaum Associates, Publishers, London.

Elman, J. L. (1993) Learning and development in neural networks: The importance of starting small. *Cognition*. Vol. 48, 71-99

Gelfand, S. A. (1998). *Hearing – an introduction to psychological and physiological acoustics*. Marcel Dekker, Inc. New York.

Greenberg, S. (1997) *On the origins of speech intelligibility in the real world*. ESCA Workshop 1997, pp. 23-32.

Greenberg, S. (2001) *From Here to Utility – Melding Phonetic Insight with Speech Technology*. Eurospeech 2001, Scandinavia.

Hermansky, H. (1990): Perceptual linear predictive (PLP) analysis of speech, *Journal of the Acoustical Society of America.* 87(4), 1738-1752.

Itahashi S. and Yokoyama S. (1976) Automatic formant extraction utilizing mel scale and equal loudness contour, *Acoustics, Speech, and Signal Processing*, IEEE International Conference on ICASSP '76

Johnson, K. (2003) *Acoustic and Auditory Phonetics,* Blackwell Publishing

Kent, R. D and Read, C. (1992) *The Acoustic Analysis of Speech*. Whurr Publishers, London.

Koponen, E., Gustavsson, L. and Lacerda, F. (2003). *Effects of Linguistic Variance on Sound-Meaning-Connections in Early Stages of Language Acquisition*. Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona.

Lindblom, B. (2000) Developmental Origins of Adult Phonology: The Interplay between Phonetic Emergents and the Evolutionary Adaptations of Sound Patterns, in *Phonetica*, Emergence and Adaptation, Diehl, R., Engstrand, O., Kingston, J. and Kohler, K. (Eds.), Vol. 57, No. 2-4.

Lippmann, R. P. (1997) Speech recognition by machines and humans. *Speech Communication*, Vol. 22.

Minsky, M. (1985) Why Intelligent Aliens will be Intelligible. In Regis, E. JR. *Extraterrestrials: Science and Alien Intelligence*. Cambridge University Press.

Ney, H. (2003) *Acoustic-Phonetic Knowledge and Statistics in Automatic Speech Recognition.* 15th International Congress of Phonetic Sciences, Barcelona

Pickles, J. O. (1988) *An Introduction to the Physiology of Hearing* Academic Press

Pinker, S. (1994) *Rules of language In Language Acquisition*. Bloom, P. (Edt.) MIT Press, Cambridge.

Plunkett, K. and Elman, J. L. (1997) *Exercises in Rethinking Innateness*, The MIT Press.

Rabiner, L. R. (1989) Proceedings of the IEEE, Vol 77, No 2.

Rumelheart, D. E. and McClelland, J. L. (1994). On learning the past tenses of English verbs. In *Language Acquisition*. Bloom, P. (Edt.) MIT Press, Cambridge.

Sonesson, B. and Sonesson, G. (1993). *Människans anatomi och fysiologi*, Almqvist & Wiksell, Falköping.

Stevens, K. N. (1998) Acoustic Phonetics, The MIT Press.

Stevens, S. S. (1957) Concerning the form of the loudness function. *Journal of the Acoustical Society of America*.

Weng, J. and Zhang, Y. (2002) *Developmental Robots: A New Paradigm*. Proc. of Second International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems. Edinburgh, Scotland

Weng, J. and Zhang, Y. (2001) *Grounded auditory development by a developmental robot*. Proc. INNS-IEEE International Joint Conference on Neural Networks, 1059-1064. Washington, DC.

Yates, G. K. (1993) The Ear as an Acoustical Transducer. *Acoustics Australia*, Vol. 21, pp. 77-81.