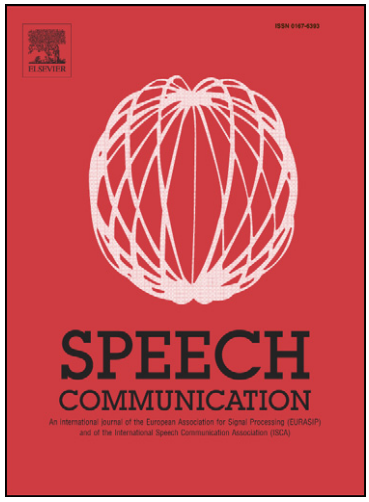
Reaching over the gap: A review of efforts to link human and automatic speech recognition research

Odette Scharenborg



Reaching over the gap: A review of efforts to link human and automatic speech recognition research

Odette Scharenborg

PII: S0167-6393(07)00010-6
DOI: 10.1016/j.specom.2007.01.009
Reference: SPECOM 1610

To appear in: *Speech Communication*

Received Date: 26 April 2006
Revised Date: 6 December 2006
Accepted Date: 18 January 2007

Please cite this article as: Scharenborg, O., Reaching over the gap: A review of efforts to link human and automatic speech recognition research, *Speech Communication* (2007), doi: 10.1016/j.specom.2007.01.009

Reaching over the gap:

A review of efforts to link human and automatic speech recognition research

Odette Scharenborg

Speech and Hearing Research Group

Department of Computer Science, University of Sheffield, Sheffield, UK

O.Scharenborg@dcs.shef.ac.uk

Address for correspondence:

Odette Scharenborg

Speech and Hearing Research Group

Department of Computer Science

Regent Court, 211 Portobello Street

Sheffield S1 4DP

UK

Telephone: (+44) 114 222 1907

Fax: (+44) 114 222 1810

E-mail: O.Scharenborg@dcs.shef.ac.uk

Abstract

The fields of human speech recognition (HSR) and automatic speech recognition (ASR) both investigate parts of the speech recognition process and have word recognition as their central issue. Although the research fields appear closely related, their aims and research methods are quite different. Despite these differences there is, however, lately a growing interest in possible cross-fertilisation. Researchers from both ASR and HSR are realising the potential benefit of looking at the research field on the other side of the 'gap'. In this paper, we provide an overview of past and present efforts to link human and automatic speech recognition research and present an overview of the literature describing the performance difference between machines and human listeners. The focus of the paper is on the mutual benefits to be derived from establishing closer collaborations and knowledge interchange between ASR and HSR. The paper ends with an argument for more and closer collaborations between researchers of ASR and HSR to further improve research in both fields.

Keywords: automatic speech recognition; human speech recognition.

1. Introduction

Both the research fields of human speech recognition (HSR) and automatic speech recognition (ASR) investigate (parts of) the speech recognition process. The two research areas are closely related since they both study the speech recognition process and the central issue of both is word recognition. However, their research objectives, their research approaches, and the way HSR and ASR systems deal with different aspects of the word recognition process differ considerably (see Section 2). In short, in HSR research, the goal is to understand how we, as listeners, recognise spoken utterances. This is often done by building computational models of HSR, which can be used for the simulation and explanation of behavioural data related to the human speech recognition process. The aim of ASR research is to build algorithms that are able to recognise the words in a speech utterance automatically, under a variety of conditions, with the least possible number of recognition errors. Much research effort in ASR has therefore been put into the improvement of, amongst others, signal representations, search methods, and the robustness of ASR systems in adverse conditions.

One might expect that in the past this common goal would have resulted in close collaborations between the two disciplines, but in reality, the opposite is true. This lack of communication is most likely to be attributed to another difference between ASR and HSR. Although both ASR and HSR claim to investigate the whole recognition process from the acoustic signal to the recognised units, an automatic speech recogniser necessarily is an end-to-end system – it must be able to recognise words from the acoustic signal – while most models of HSR only cover parts of the human speech recognition process (Nearey, 2001; Moore & Cutler, 2001). Furthermore, in ASR, the algorithms and the way to train the ASR systems are completely understood from a mathematical point of view, but in practice it has so far proved impossible to get the details sufficiently right to achieve a recognition

performance that is even close to human performance. Human listeners, on the other hand, achieve superior performance, but many of the details of the internal processes are unknown.

Despite this gap that separates the two research fields, there is a growing interest in possible cross-fertilisation (Furui, 2001; Hermansky, 2001; Huckvale, 1998; Kirchhoff & Schimmel, 2005; Moore, 1995; Moore & Cutler, 2001; Pols, 1999; Scharenborg, 2005a, 2005b; Scharenborg et al., 2005; ten Bosch, 2001). (For a historical background on the emergence of this gap, see Huckvale (1998).) This is, for instance, clearly illustrated by the organisation of the workshop on *Speech recognition as pattern classification* (July 11-13, 2001, Nijmegen, The Netherlands), the organisation of the special session *Bridging the gap between human and automatic speech processing* at *Interspeech 2005* (September 6, 2005, Lisbon, Portugal) and of course with the coming about of this 'Speech Communication' special issue on *Bridging the gap between human and automatic speech processing*.

It is generally acknowledged within the ASR community that the improvement in ASR performance observed in the last few years can to a large extent be attributed to an increase in computing power and the availability of more speech material to train the ASR systems (e.g., Bourlard et al., 1996; Moore & Cutler, 2001). However, the incremental performance is asymptoting to a level that falls short of human performance. It is to be expected that further increasing the amount of training data for ASR systems will not result in recognition performances that are even approaching the level of human performance (Moore, 2001, 2003; Moore & Cutler, 2001). What seems to be needed is a change in approach – even if this means an initial worsening of the recognition performance (Bourlard et al., 1996). As pointed out by Moore and Cutler (2001): "true ASR progress is not only dependent on the analysis of ever more data, but on the development of more structured models which better exploit the information available in existing data". ASR engineers hope to get clues about those "structured models" from the results of research in HSR. Thus, from the point of view of

ASR, there is hope of improving ASR performance by incorporating essential knowledge about HSR into current ASR systems (Carpenter, 1999; Dusan & Rabiner, 2005; Furui, 2001; Hermansky, 1998; Maier & Moore, 2005; Moore, 2003; Moore & Cutler, 2001; Pols, 1999; Scharenborg et al., 2007; Strik, 2003, 2006; Wright, 2006).

With respect to the field of HSR, specific strands in HSR research hope to deploy ASR approaches to integrate partial modules into a convincing end-to-end model (Nearey, 2001). As pointed out above, computational models of HSR only model parts of the human speech recognition process; an integral model covering all stages of the human speech recognition process does not yet exist. The most conspicuous part of the recognition process that, until recently, virtually all models of human speech recognition took for granted is a module that converts the acoustic signal into some kind of symbolic segmental representation. So, unlike ASR systems, most existing HSR models cannot recognise real speech, because they do not take the acoustic signal as their starting point. In 2001, Nearey pointed out that the only working models of lexical access that take an acoustic signal as input were ASR systems. Mainstream ASR systems however are usually implementations of a specific computational paradigm and their representations and processes need not be psychologically plausible. In their computational analysis of the HSR and ASR speech recognition process, Scharenborg et al. (2005) showed that some ASR algorithms serve the same functions as analogous HSR mechanisms. Thus despite first appearances, this makes it possible to use certain ASR algorithms and techniques in order to build and test more complete computational models of HSR (Roy & Pentland, 2002; Scharenborg, 2005b; Scharenborg et al., 2003, 2005; Wade et al., 2001; Yu et al., 2005).

Furthermore, within the field of ASR there are many (automatically or hand-annotated) speech and language corpora. These corpora can easily and quickly be analysed using ASR systems, since ASR systems and tools are able to process large amounts of data in

a (relatively) short time. This makes ASR techniques valuable tools for the analysis and selection of speech samples for behavioural experiments and the modelling of human speech recognition (de Boer & Kuhl, 2003; Kirchhoff & Schimmel, 2005; Pols, 1999).

Researchers from both ASR and HSR are thus realising the potential benefit of looking at the research field on the other side of the gap. This paper intends to give a comprehensive overview of past and present efforts to link human and automatic speech recognition research. The focus of the paper is on the mutual benefits to be derived from establishing closer collaborations and knowledge interchange between ASR and HSR. First a brief overview of the goals and research approaches in HSR and ASR is given in Section 2. In Section 3, we discuss the performance difference between machines and human listeners for various recognition tasks and what can be learned from comparing those recognition performances. Section 4 discusses approaches for improving the computational modelling of HSR by using ASR techniques. Section 5 presents an overview of the research efforts aimed at using knowledge from HSR to improve ASR recognition performance. The paper ends with a discussion and concluding remarks.

2. Human and automatic speech recognition

In this section, the research fields and approaches of human speech recognition (Section 2.1) and automatic speech recognition (Section 2.2) will be discussed briefly. A more detailed explanation of the two research fields would be beyond the scope of this article; the reader is referred to textbook accounts such as Harley (2001) for an in-depth coverage of the research field of human speech recognition, and to textbook accounts such as Rabiner and Juang (1993) or Holmes and Holmes (2002), for an explanation of the principles of automatic speech recognition. Comprehensive comparisons of the research goals and approaches of the two fields (Huckvale, 1998; Moore & Cutler, 2001; Scharenborg et al., 2005), as well as comparisons of the computational functioning and architectures (Scharenborg et al., 2005) of

the speech recognition process in automatic recognition systems and human listeners can also be found in the literature. Furthermore, Dusan and Rabiner (2005) provide an extensive comparison between human and automatic speech recognition along six key dimensions (including the architecture of the speech recognition system) of ASR.

2.1. Human speech recognition

To investigate the properties underlying the human speech recognition process, HSR experiments with human subjects are usually carried out in a laboratory environment. Subjects are asked to carry out various tasks, such as:

- *Auditory lexical decision:* Spoken words and non-words are presented in random order to a listener, who is asked to identify the presented items as a word or a non-word.
- *Phonetic categorisation:* Identification of unambiguous and ambiguous speech sounds on a continuum between two phonemes.
- *Sequence monitoring:* Detection of a target sequence (larger than a phoneme, smaller than a word), which may be embedded in a sentence or list of words/non-words, or in a single word or non-word.
- *Gating:* A word is presented in segments of increasing duration and subjects are asked to identify the word being presented and to give a confidence rating after each segment.

In these experiments, various measurements are taken, such as reaction time, error rates, identification rates, and phoneme response probabilities. Based on these measurements, theories about specific parts of the human speech recognition system are developed. To put the theories to further test, they are implemented in the form of computational models. Being implementations of a theory, computational models are regarded as proof of principle of a theory (Norris, 2005). However, to actually implement a computational model, several assumptions have to be made, for instance about the nature of the input. It is of the utmost importance that these assumptions are chosen appropriately, such that the resulting

computational model is able to model the observed human data, on the one hand, and is psychologically (and biologically) plausible, on the other (Norris, 2005; Tuller, 2003). The chosen implementation of the computational model should provide an affirmative answer to the question whether speech recognition can really work like this. Additionally, a good computational model should also be able to make accurate predictions of aspects of the phenomenon under investigation that do not directly follow from the observed data or the literature (Norris, 2005; Tuller, 2003). These predictions can then be used for the definition of new behavioural studies and subsequently, if necessary, a redefinition or adaptation of the original theory on which the computational model was based. Various models of HSR (e.g., Luce et al., 2000; Marslen-Wilson, 1987; McClelland & Elman, 1986; Norris, 1994) have been developed that are capable of simulating data from behavioural experiments.

Most data on human *word* recognition involve measures of how quickly or accurately words can be identified. A central requirement of any model of human word recognition is therefore that it should be able to provide a continuous measure (usually referred to as 'activation' or 'word activation') associated with the strength of different lexical hypotheses over time. During the human speech recognition process word hypotheses that overlap in time compete with each other. This process is referred to as (lexical) competition. Virtually all existing models of HSR assume that the activation of a word hypothesis at a certain point in time is based on its initial activation (due to the information in the acoustic signal and prior probability) and the inhibition caused by other activated words. The word activation score, then, can be compared to the performance of listeners in experiments where they are required to make word-based decisions (such as the above-described auditory lexical decision experiments).

The investigation into how human listeners recognise speech sounds and words from an acoustic signal has a long history. Miller and Nicely's famous 1954 paper has led to many

investigations into the processing of speech by the auditory system analysing sound confusions made by human listeners as a function of signal-to-noise ratios (see Allen, 2004, and references therein). A detailed account of how listeners recognise spoken words is provided by McQueen (2004). In short, there are two major theories of human speech recognition, which are presented here as extreme standpoints. The first theory, referred to as 'episodic' or 'sub-symbolic' theory, assumes that each lexical unit is associated with a large number of stored acoustic representations (e.g., Goldinger, 1998; Klatt, 1979, 1989). During the speech recognition process, the incoming speech signal is compared with the stored acoustic representation. The competition process will decide which representation (and thus word) is recognised. On the other hand, 'symbolic' theories of human speech recognition hold that human listeners first map the incoming acoustic signal onto prelexical representations, e.g., in the form of phonemes, after which the prelexical representations are mapped onto the lexical representations stored in the form of a sequence of prelexical units (e.g., Gaskell & Marslen-Wilson, 1997; Luce et al., 2000; McClelland & Elman, 1986; Norris, 1994). The speech recognition process in symbolic theories thus consists of two levels: the prelexical level and the lexical level, at which the competition process takes place.

2.2. Automatic speech recognition

The development of an ASR system consists of four essential steps: feature extraction, the acoustic modelling, the construction of a language model, and the search. First, in the so-called front-end, numerical representations of speech information, or features, are extracted from the raw speech signal. These features provide a relatively robust and compact description of the speech signal, ideally, preserving all information that is relevant for the automatic recognition of speech. These features describe spectral characteristics such as the component frequencies found in the acoustic input and their energy levels. Second, in the acoustic modelling stage, an acoustic model is created for each recognition unit (also known

as sub-word units, e.g., phones). In most current state-of-the-art ASR systems, the acoustic models are based on the hidden Markov Model (HMM) paradigm (see for an introduction, Rabiner & Juang (1993)). HMMs model the expected variation in the signal statistically. Probability density functions for each sub-word HMM are estimated over all acoustic tokens of the recognition unit in the training material.

ASR systems use language models to guide the search for the correct word (sequence). Most ASR systems use statistical *N*-gram language models which predict the likelihood of a word given the *N* preceding words. The a priori probabilities are learned from the occurrences and co-occurrences of words in a training corpus. An ASR system can only recognise those words which are present in its lexicon. Each word in the lexicon is built from a limited number of sub-word units. The type of unit used to describe the words in the lexicon is identical to the type of unit represented by the acoustic models. So, if the acoustic models represent phonemes, the units used to describe the words in the lexicon are also phonemes.

During word recognition, then, the features representing the acoustic signal are matched with the succession of acoustic models associated with the words in the internal lexicon. Most ASR systems use an integrated search: all information (from the acoustic model set, lexicon, and language model) is used at the same time. However, lately, also multi-stage ASR systems are developed, in which a first stage recogniser converts the acoustic signal into an intermediate probabilistic representation (of, for instance, phones) after which the second stage recogniser maps the intermediate representation onto lexical representations. The advantage of such multi-stage recognition systems is that in a second (or subsequent) recognition step more detailed information can be used, for instance by integrating more powerful language models into the system. During recognition, the likelihood of a number of hypothesised word sequences (paths) through the complete search space is computed (using Bayes' Rule), and then a trace back is performed to identify the words that were recognised

on the basis of the hypothesis with the highest score at the end of the utterance (or the hypothesis with the highest score after a number of recognised words, depending on the length of the influence of the language model). In addition to the best-scoring word (sequence), a word graph can be constructed, which is a compact and efficient representation for storing an *N*-best list. It contains those path hypotheses whose scores are closest to the best scoring path.

It is important to point out that a standard ASR system is not capable of deciding that a stream of feature vectors belongs to a non-word: an ASR system will always come up with a solution in terms of the items that are available in the lexicon. (It is possible to configure an ASR system such that it rejects inputs if the quality of the match with the words in the vocabulary is below a certain minimum. However, this is not the same as detecting that the input speech contains non-words.). Furthermore, an ASR system can only recognise what it has observed in the training material. This means that an ASR system is able to recognise words that are not present in the training material as long as the new word's sub-word units are known to the ASR system and the new word has been added to the recogniser's lexicon; once the new word contains a sub-word unit not observed in the training material, the ASR system will not be able to recognise it. Lastly, ASR systems are usually evaluated in terms of accuracy, the percentage of the input utterances that is recognised correctly, or in terms of word error rate (WER): the number of word insertions, word deletions, and word substitutions divided by the total number of words.

3. The difference in human and machine speech recognition performance

Over the past decades the recognition performance of ASR systems has drastically increased. The question now is how close the performance of ASR systems is to the best possible recognition performance on a given task. It is generally assumed that human listeners outperform ASR systems on most tasks (e.g., Cutler & Robinson, 1992; Moore & Cutler,

2001); therefore, several studies have been carried out in which machine recognition performance is compared to human performance to investigate the size of the performance difference. Secondly, comparative studies have been carried out to investigate what it is that makes human speech recognition so superior to machine speech recognition, and what can be learned from human speech recognition to improve ASR performance.

A problem that arises when one wants to make a comparison between human and machine recognition performance is that the performance of ASR systems is usually measured in terms of (word) accuracy, while HSR researchers are not only interested in the number of correct and incorrect responses of a subject but also, for instance, in the (relative) speed (or 'reaction time', which gives an indication of the relative difficulty of the recognition task: a slower reaction time indicates that it takes more time for a word's activation to become high enough to be recognised) with which a subject fulfils a certain task. Reaction time as measured in HSR experiments is hardly an issue in ASR applications, the latter requires adding (perhaps complex) features to existing ASR systems to make it possible to assess reaction times for machines; therefore, most comparative studies of human-machine performance describe the performances in terms of word accuracy or error rates. An exception is a feasibility study by Cutler and Robinson (1992) who compared human and machine recognition performance by using reaction time as a metric, showing a significant positive correlation between human and machine reaction times for consonants, but not for vowels. Such a comparison of the performance of human listeners and machines using reaction times can highlight differences in the processing of, for instance, phonemes and vowels.

The recognition performances in terms of accuracy or error rates of human listeners and machines have been compared at several levels: words (Carey & Quang, 2005; Lippmann, 1997; van Leeuwen et al., 1995), phonemes (Cutler & Robinson, 1992; Meyer et al., 2006; Sroka & Braida, 2005), and (articulatory) features (Cooke, 2006; Meyer et al., 2006;

Sroka & Braida, 2005); for different listening conditions (clean speech: Cutler & Robinson, 1992; Lippmann, 1997; Meyer et al., 2006; van Leeuwen et al., 1995; noisy speech: Carey & Quang, 2005; Cooke, 2006; Lippmann, 1997; Meyer et al., 2006; Sroka & Braida, 2005; degraded speech: Sroka & Braida, 2005); and with respect to amount of training material (Moore, 2001, 2003, Moore and Cutler 2001). In his comprehensive – often cited – study, Lippmann (1997) compared the performance of human listeners and the best performing ASR system (for each speech corpus) on six word recognition tasks comprising varying speaking styles and lexicon sizes. The recognition tasks varied from relatively easy, such as the recognition of digits spoken in isolation and in short sequences (the TI-digits corpus; Leonard, 1984) and the recognition of letters pronounced in isolation (the alphabet letters corpus; Cole et al., 1990) to far more difficult tasks, such as the recognition of read sentences from the *Wall Street Journal* (the North American Business News corpus; Paul & Baker, 1992) and the recognition of spontaneous telephone conversations (the NIST Switchboard corpus; LDC, 1995). Except for two speech recognition tasks, the human and machine performances were measured on materials from the same corpora. Van Leeuwen et al. (1995) performed a study, in which they compared the performance of 20 native speakers of English, 10 non-native speakers of English, and 3 ASR systems on 80 sentences from the Wall Street Journal database. Both studies showed that the word error rates obtained by ASR systems are significantly higher than those obtained by human listeners – often one or more orders of magnitude (for instance, human listeners obtained an error rate of 4% on data taken from Switchboard, while the best ASR systems obtained an error rate of 43% (Lippmann, 1997)); even non-native speakers significantly outperformed ASR systems. At the phoneme level, the recognition results are in line with those found at the word level: machine error rates were significantly higher than human error rates (Cutler & Robinson, 1992; Meyer et al., 2006). The type of errors made by the machine, however, are more or less similar (but higher) to the

errors made by the human listeners, suggesting that the human listeners and the machine had similar type of recognition difficulties (Cutler & Robinson, 1992).

A more detailed analysis of the errors made by human listeners and ASR systems revealed that although human listeners and ASR both find content words more difficult to recognise correctly than function words, the types of substitutions made in content word errors are different: human listeners make fewer inflection errors than machines. One can suspect that inflection errors are less severe than the substitution of a word with any given (acoustically similar) other word, because it will not impair the understanding of an utterance (van Leeuwen et al., 1995). This is a point where ASR systems need to be improved, perhaps at the level of language modelling.

In *adverse* listening conditions (such as situations where the signal to noise ratio is very low), the difference in human-machine recognition performance increases even further (Carey & Quang, 2005; Cooke, 2006; Lippmann, 1997; Meyer et al., 2006; Sroka & Braida, 2005). It is not just that humans are better able than machines to recognise speech in adverse listening conditions, but they are also better at recognising speech when the background noise is non-stationary (e.g., noise with a speech-like spectrum and modulation spectrum; Carey & Quang, 2005). However, one needs to take into account that human listeners have multiple sources of information available that are unavailable to the ASR system. A human listener is able to use knowledge about the world, the environment, the topic of discourse, etc. This 'higher-level' knowledge is not incorporated in the statistical language models used in ASR systems. When comparing human and machine recognition performance on a task where the higher-level information flows are removed, i.e., the recognition of nonsense words and sentences, human recognition performance is however still much better than machine recognition performance (Lippmann, 1987; Meyer et al., 2006). One speech recognition corpus that has specifically been designed in order to perform unbiased tests between human

and machine performance and which satisfies requirements for both ASR and HSR tests is the OLdenburg LOgatome (OLLO) speech corpus (Wesker et al., 2005): a corpus consisting of phonemes embedded in logatomes, i.e., three-phoneme sequences (CVC and VCV) with no semantic information.

These differences in machine and human performance, even in tasks where there is no higher-level information to help human speech recognition, suggest that human listeners and ASR systems use different acoustic cues during speech recognition. This is further backed-up by 'lower-level' differences between human listeners and ASR systems: human listeners are able to use all information that is present in the acoustic signal to differentiate between phones and thus words, while ASR systems can only use the information that is encoded in the acoustic features (e.g., Mel-frequency cepstral coefficients or perceptual linear prediction features); ASR systems thus do not have available all information that is available to human listeners. Knowledge about which cues are present in the speech signal and are most robustly detected by human listeners (usually tested in adverse listening conditions; e.g., Cooke, 2006; Sroka & Braida, 2005) can be used to improve machine recognition performance, more specifically the feature extraction in the front-end of an automatic speech recogniser. Furthermore, it will help to direct one's recognition improvement efforts only to those phonemes that are often misrecognised, instead of trying to improve the complete phoneme set, and thus trying to improve phonemes that are already recognised quite well. Phonological feature analyses carried out on the human and machine speech recognition results can help identify those features (and phonemes) that are frequently recognised correctly and incorrectly, and thus which cues are being used by human listeners and not by ASR systems and vice versa. Automatic recognition systems outperform human listeners in the identification of *plosives* and *non-sibilant fricatives* (Cooke, 2006), while the recognition of articulation of *place* is similar for humans and machines (Cooke, 2006; Sroka & Braida, 2005

(although Meyer et al. (2006) report a poorer performance for the ASR system)). It is thus needless to try to improve the distinction, for instance, between different plosive consonants. On the other hand, *voicing* information (Cooke, 2006; Meyer et al., 2006; Sroka & Braida, 2005) is recognised much more poorly by machines than by human listeners. In order to improve the ASR's recognition performance, it is thus useful, even necessary, to improve the 'recognition' of *voicing*.

4. HSR: Using knowledge from ASR

4.1. Computational modelling of human speech recognition

Computational models of HSR deal only with particular components of the speech recognition system, and many parts of the speech recognition system remain unspecified. This makes it difficult to assess whether the assumptions underlying the computational model are actually consistent with an effective complete recognition system. One important contribution of the ASR community to the field of HSR is to provide the techniques to implement more complete computational models, for instance, models that can recognise real speech instead of using hand-made segmental transcriptions or other artificial forms of input (Roy & Pentland, 2002; Scharenborg, 2005b; Scharenborg et al., 2003, 2005; Wade et al., 2001; Yu et al., 2005). For instance, removing the need for hand-made transcriptions will make it possible to test the computational model on the same speech material as was used for the behavioural studies, thus removing any dissimilarity between the input data of the computational model and the human listener. Additionally, HSR modelling has tended to avoid detailed analysis of the problems of robust speech recognition given real speech input (Scharenborg et al., 2005). ASR on the other hand has to deal with those problems in order to get a system recognising speech at a reasonable performance level. Thus, if the speech-driven computational model is able to correctly simulate the observed human behaviour, this will strengthen the theory and

the assumptions on which the model is based, since some of its underlying assumptions (e.g., the one referring to the input representation) have been removed. It will, however, not prove that the theory is correct; though, if a computational implementation of a certain theory is not able to correctly model the data, this will seriously diminish the likelihood of the correctness of the theory on which the model is based. ASR techniques can thus be used to build more complete computational models of HSR. In doing so, the simulation and explanation power of those computational models will increase.

Computational models of lexical access operating on real speech have been developed that have proven to be able to simulate data found in behavioural studies. Wade et al. (2001) successfully used the MINERVA2 memory model (Hintzman, 1984, based on the principles underlying the episodic theory (see Section 2.1)) to replicate effects of word category frequency on the recognition of related and unrelated words observed in human word recognition. Scharenborg et al. (2003, 2005) built a computational model on the basis of the (symbolic) theory underlying the *Shortlist* model (Norris, 1994) using techniques from ASR, which was able to simulate well-known phenomena from three psycholinguistic studies on human word recognition. Like computational models of lexical access, computational models of word acquisition by infants have benefited from the use of ASR techniques. Recently, computational models have been developed that explore issues of early language learning using methods of computer vision and ASR (Roy & Pentland, 2002; Yu et al., 2005). On the basis of raw (infant-directed) speech material and video material (consisting of video images of single objects), the computational models simulate parts of the process of language learning in infants: speech segmentation, word discovery, and visual categorisation. The models proved to be reasonably successful in segmenting the speech, finding the words, and making pairs of audio-stimuli and visual stimuli; i.e., associating visual objects with the correct input speech fragments.

Given that these computational models of lexical access (Scharenborg, 2005b; Scharenborg et al., 2003, 2005; Wade et al., 2001) and word acquisition (Roy & Pentland, 2002; Yu et al., 2005) operate on acoustic signals, it is to be expected that their recognition performance is degraded compared to computational models which process human-generated representations of the speech signal (which have recognition performances close to 100% correct). Scharenborg et al. (2005) used their computational model of lexical access for the automatic recognition of words produced in isolation. The SpeM model was able to correctly recognise 72.1% of the words. However, despite the degraded recognition results, SpeM was able to correctly simulate human speech recognition data. It is to be expected that this gap in recognition performance between speech-driven computational models and computational models which process human-generated representation of the speech signal will diminish in the future, since it is likely that the recognition performance of the speech-driven computational models will improve, for instance, by using more robust signal representations. Speech-driven computational models are thus useful tools that will lead to an improved understanding of the human speech recognition process and better theories. Furthermore, since speech-driven computational models can be tested using the same stimuli as human listeners and perform at reasonable levels (with the expectation that the performance will further improve), these computational models can also be used to investigate the effect of changing specific variables instead of having to carry out expensive and time-consuming behavioural studies.

Automatic speech processing techniques have also been used to investigate the properties underlying infant-directed and adult-directed speech (de Boer & Kuhl, 2003; Kirchhoff & Schimmel, 2005). Infant-directed speech is characterised by greater distances between the means of vowel classes measured in formant space and greater variances than adult-directed speech (e.g., de Boer & Kuhl, 2003). The question that arises is why speech

produced by parents while talking to their infants differs so much from their speech when talking to another adult. Using ASR techniques, the hypothesis that infant-directed speech is easier to learn has been investigated (de Boer & Kuhl, 2003) and proven to be correct: they showed that their learning algorithm learned the positions of vowels in the acoustic space more accurately on the basis of infant-directed speech than on the basis of adult-directed speech. Besides their role in building computational models of human speech recognition, ASR techniques thus also play an important role in the analysis and answering of developmental and cognitive questions (de Boer & Kuhl, 2003).

A third contribution of ASR techniques for the computational modelling of human speech recognition is in the context of speech recognition in 'noisy' listening conditions. In everyday listening conditions, the speech signal that reaches our ears is hardly ever 'clean'. Usually, the speech signal is embedded in a mixture of other sound sources, frequently alongside additional energy reflected from reverberant surfaces. The listener (whether human or machine) is thus faced with the task of separating out the acoustic input into individual sources. This process is known as *auditory scene analysis* (Bregman, 1990) and has been investigated for several decades. In this research field, there has been a steady growth in computational models trying to model human recognition, for instance, in listening conditions where a stronger signal masks a weaker one within a critical band or with a low signal-to-noise ratio (Cooke, 2006). For the modelling of human recognition behaviour in such degraded listening conditions, ASR systems that make use of a missing data strategy (e.g., Cooke et al., 2001) have been successfully applied. Cooke and Ellis (2001) and Wang and Brown (2006) provide a comprehensive explanation of the field of auditory scene analysis and present an overview of the existing computational models of human recognition in noisy environments.

4.2. Fine-phonetic detail and computational modelling of human speech recognition

There is now considerable evidence from psycholinguistic and phonetic research that sub-segmental (i.e., subtle, fine-grained, acoustic-phonetic) and supra-segmental (i.e., prosodic) detail in the speech signal modulates human speech recognition, and helps the listener segment a speech signal into syllables and words (e.g., Davis et al., 2002; Kemps et al., 2005; Salverda et al., 2003). It is this kind of information that appears to help the human perceptual system distinguish short words (like *ham*) from the longer words in which they are embedded (like *hamster*). However, currently no computational models of HSR exist that are able to model the contributions of this fine-phonetic detail (Hawkins, 2003).

Scharenborg et al. (2006) present a preliminary study of the effectiveness of using articulatory features (AFs) to capture and use fine-grained acoustic-phonetic variation during speech recognition in an existing computational model of human word recognition. Many more studies into the automatic recognition of AFs have been carried out (e.g., King & Taylor, 2000; Kirchhoff, 1999; Livescu et al., 2003; Wester, 2003, Wester et al., 2001). AFs describe properties of speech production and can be used to represent the acoustic signal in a compact manner. AFs are abstract classes which characterise the most essential aspects of articulatory properties of speech sounds in a quantised form leading to an intermediate representation between the signal and the lexical units (Kirchhoff, 1999). The AFs often used in AF-based ASR systems are based on features proposed by Chomsky and Halle (1968), e.g., voice, nasality, roundedness, etc.. In the field of ASR, AFs are often put forward as a more flexible alternative (Kirchhoff, 1999; Wester, 2003; Wester et al., 2001) to modelling the variation in speech using the standard 'beads-on-a-string' paradigm (Ostendorf, 1999), in which the acoustic signal is described in terms of (linear sequences of) phones, and words as phone sequences.

5. ASR: Using knowledge from HSR

There are several areas in the field of human speech recognition that might benefit ASR research, such as experimental and neuroimaging studies of human speech perception. In this section, we will discuss approaches and techniques based on knowledge derived from HSR research that have been applied for the improvement of ASR systems. Some approaches and techniques are well-established, while others are new and the future will show whether they will bring the desired improvements.

Some of the properties of human speech recognition that are well accepted and have proven useful in ASR come from the field of human auditory recognition. The acoustic features extracted from the speech signal by the acoustic pre-processor usually take into account the fact that in human hearing the spectral resolution is better at low frequencies than at high ones. Dominant ASR feature extraction techniques, then, are based on frequency warped short-term spectra of speech where the spectral resolution is better at low frequencies than at high ones (e.g., Mel Frequency Cepstral Coefficients (Davis & Mermelstein, 1980), Perceptual Linear Predictive (Hermansky, 1998, 2001)).

Although these acoustic features have proven to be rather successful, the difference between human and machine performance suggests that there is still information in the speech signal not being extracted or at least not being used for speech recognition by ASR systems. For instance, phase spectrum information is discarded in the extraction of the acoustic features; while of course human listeners do have this information available. In a series of experiments over a range of signal to noise ratios Alsteris and Paliwal (2006) have shown that when human listeners are presented with re-constructed speech that does not contain phase information, the intelligibility is far worse than when the original signal is presented (thus including the phase information). This suggests that there might be important – maybe even essential – information in the phase spectrum for the recognition of speech. Thus, an acoustic

feature set that also represents information from the phase spectrum may result in improved ASR performance. An additional source of information on missing or unused information in acoustic features comes from a different, but related, research area: studies of simulations of cochlear implant speech processing (e.g., Shannon et al., 1995; Qin & Oxenham, 2003; and references therein) and studies of the intelligibility of different types of speech for hearing-impaired and normal-hearing listeners under a variety of listening conditions (Krause & Braida, 2002, and references therein) can be used to find the acoustic information in the speech signal that is important for human speech recognition but which is currently not extracted from the speech signal or not being used during the automatic recognition of speech.

As pointed out above, human listeners greatly outperform machines in all speech recognition tasks. Several researchers (Cooke et al., 2001; Hermansky, 1998; Lippmann 1997) have, therefore, argued that the search for ASR systems that perform robustly in adverse conditions have much to gain by examining the basis for speech recognition in listeners. Human listeners are capable of dealing with missing (because of the presence of noise, of errors or bandlimiting in the transmission channel) speech information. ASR systems have been built that acknowledge the fact that some parts of the speech signal are masked by noise. 'Noisy' speech fragments are not simply thrown away; instead, information in noisy regions is used by these systems to define an upper bound on the energy present in any of the constituent sources at the time-frequency location. For instance, if one was hypothesising a /s/ but the energy in a high frequency but unreliable region was too low, that hypothesis would be (probabilistically) weighted against (Barker et al., 2005; Cooke, 2006; Cooke et al., 1994, 2001; Raj et al., 2004; Seltzer et al., 2004; Srinivasan & Wang, 2005).

Although HMMs have proven to be powerful tools for the automatic recognition of speech, the asymptoting recognition accuracies and the fact that HMMs have short-comings (e.g., the first-order Markov assumption, that the probability of a certain observation at time *t*

only depends on the observation at time *t*-1, is incorrect for natural speech) brought suggestions for a new research approach that moves (partly) away from HMM-based approaches in the direction of template-based approaches (Axelrod & Mason, 2004; De Wachter et al., 2003; Maier & Moore, 2005; Strik, 2003, 2006; Wade et al., 2002), which is very similar to the episodic theory of human speech recognition (Goldinger, 1998). Template-based speech recognition relies on the storage of multiple templates. During speech recognition, the incoming speech signal is compared with the stored templates, and the template with the smallest distance to the input serves as the recognition output. Several methods have been used for matching the incoming speech with the stored templates, such as the memory model MINERVA2 (Maier & Moore, 2005; Wade et al., 2002) and techniques such as dynamic time-warping (Axelrod & Mason, 2004; De Wachter et al., 2003). On small tasks, such as digit recognition, these methods proved successful. The future will show whether these approaches can be extended such that they will be useful for large vocabulary recognition tasks.

Another approach for dealing with the fact that more training data will not provide higher levels of recognition performance is by using 'better' training data (Kirchhoff & Schimmel, 2005). Studies into child language acquisition have shown that so-called 'motherese' or 'parentese' speech (i.e., speech produced by an adult when talking to an infant or child) helps the infant in language acquisition (de Boer & Kuhl, 2003). Studies have shown that the greater distance between the means of the vowel classes in infant-directed speech might facilitate phonetic category learning. However, when infant-directed speech was used to train an ASR system, the performance on adult-directed speech was lower than when the system was trained on adult-directed speech (Kirchhoff & Schimmel, 2005). In a way, this was to be expected, because of the problem with training-test mismatch in all tasks involving probabilistic pattern recognition. This can be taken as a hint that HSR is not just probabilistic

pattern recognition, and that we must search for different solutions of the pervasive problem that the input signals that we must recognise always differ to some extent from everything perceived in the past. One possible approach is the memory-prediction theory (Hawkins, 2004).

In addition to the training data for the acoustic models, a second important data source for the training of ASR systems is the data to train the language model. As pointed out above, a huge difference between human and machine speech recognition is that humans are able to use contextual (higher-order) information to 'guide' the recognition process, while ASR systems use paradigms based on statistics without higher-order information. This contextual information for human listeners is not just restricted to word frequency and/or the probability of co-occurrence of the current and the previous word (e.g., Marslen-Wilson, 1987). To improve the language modelling of ASR systems, contextual information should be incorporated (Carpenter, 1999; Scharenborg, 2005a; ten Bosch, 2001). Contextual information could, for instance, be integrated into ASR systems using priming-like processes, in which the recognition of a word boosts the a priori probability of another word, as is found for humans.

Human listeners are able to recognise a word before its acoustic realisation is complete. Contemporary ASR systems, however, compute the likelihood of a number of hypothesised word sequences, and identify the words that were recognised on the basis of a trace back of the hypothesis with the highest eventual score, in order to maximise efficiency and performance. The fact that ASR systems are not able to recognise words before their acoustic offset (or actually, do not 'want' to because of performance issues) prolongs the response time in a dialogue system. In order to build a dialogue system that is capable of responding in a 'natural' way (thus within the time a human being would make a response), it is necessary that the automatic recognition process is speeded up. A suggestion is to build an ASR system that is able to recognise a word before its acoustic offset (Carpenter, 1999;

Dusan & Rabiner, 2005; Scharenborg et al. (2007)). Scharenborg et al. (2007), actually implemented such a system on the basis of a computational model of human speech recognition. The first results are promising. However, it is also possible that the absence of between-turn latencies in human-human dialogues is due to the capability of predicting what the interlocutor is going to say and when (s)he will finish speaking (Tanaka, 2000). If this turns out to be the case, it would be ill-advised to try and tackle the issue on the level of the ASR module, rather than on the dialogue management level, in human-system interaction.

6. Discussion and concluding remarks

Research has shown that the difference between human and machine recognition performance is an undeniable fact of present day life: human listeners outperform ASR systems by one or more orders of magnitude on various recognition tasks. Human listeners are far better at dealing with accents, noisy environments, differences in speaking style, speaking rate, etc.. In short, they are much more flexible than ASR systems (Pols, 1999). ASR systems are generally trained for a specific task and listening environment; thus, for one accent (usually the standard accent/pronunciation), telephone speech, one speaking rate, one speaking style, a certain lexicon, in a noisy or noiseless environment. And they tend to perform rather well on these tasks. However, once, for instance, the speech style or type or listening environment changes (e.g., different accent or speaking rate, different kind of background noise), the performance of the ASR system will typically deteriorate dramatically. To accommodate for these changes, an ASR system usually needs to be retrained. Despite the fact that there are mechanisms to adapt ASR systems dynamically (see for an overview, Moore & Cunningham, 2005), human listeners adapt remarkably faster and better to any change than current ASR systems. It is for this superior flexibility and recognition performance that researchers from the field of ASR are looking at the research field on the other side of the gap: they are looking for knowledge about human speech recognition that they can use to improve machine recognition

performance. Likewise, researchers on human speech recognition are looking at the research field on the other side of the gap in order to find the tools and techniques they can use to analyse their data, test their hypotheses, and build improved and more complete end-to-end computational models of human speech recognition.

However, integrating knowledge from human speech recognition into ASR systems is not a trivial enterprise. First of all, there is still little known about the details of human speech recognition. Secondly, of the details of human speech recognition that are understood, it is unclear which properties are relevant for the improvement of ASR systems. Using the wrong knowledge might actually be counterproductive. After deciding which properties of human speech recognition could be beneficial, a new problem arises: how should this knowledge be integrated into an ASR system? It is far from straightforward to implement properties or knowledge about (partial) theories of human speech recognition into an ASR system. This explains the limited progress that has been made in integrating knowledge from HSR into ASR systems. From the side of HSR, first attempts to integrate ASR techniques with computational models have proven to be successful, but the recognition performance of the computational models still falls short of human recognition performance. For these computational models to become really useful for the simulation and prediction of human behaviour, the recognition performance has to go up. This means that computational models that operate on speech should be including more advanced ASR techniques than they have been using so far.

Communication between researchers from the field of ASR, generally engineers, and researchers from the field of HSR, generally psycholinguists, is not easy. Although they use the same concepts, they use a different vocabulary: ASR uses terminology such as dynamic programming and pre-processing, whereas HSR is described in terms of lexical competition and auditory recognition. Furthermore, identical words sometimes have (slightly) different

meanings in the two fields. These difficulties not only exist between researchers from ASR and HSR, they exist everywhere where people from different research backgrounds start working together (see Voskuhl (2004) for a case-study between ASR engineers and auditory modelling researchers). Nevertheless, these difficulties should not stop us. If both fields want to continue to make progress, a multi-disciplinary dialogue is mandatory; there is still a lot to learn from the research field on the 'other side of the gap'. For ASR systems to be able to make use of knowledge on human speech recognition, ASR researchers need a better understanding of the theories of human speech recognition; this can only be achieved by communicating with researchers from the field of HSR. Furthermore, ASR researchers should explain to HSR researchers what they would like to know about human speech recognition in order to make their ASR systems more robust to the variability in speaker, speaking style, environment, etc. Likewise, a multi-disciplinary dialogue is of great importance to advance the creation of computational models that move beyond the implementation of partial theories of human speech recognition. HSR researchers should explain to ASR researchers which parts in their (partial) theories they have trouble implementing or dealing with. In short, both research fields will benefit tremendously if researchers of human and automatic speech recognition work together at defining the research questions (in both fields). More experimental work with human listeners to understand how humans adapt to talker and environmental variability is definitely needed.

As is reviewed in Section 3, already a fair number of human-machine performance comparisons have been carried out. Nevertheless, there is still much to be learned from comparing human and machine performance; for instance, to identify which parts of human speech recognition knowledge can be used for the improvement of ASR system, and where improvement for the ASR system is still to be gained. Research into the differences in availability of acoustic cues has already shown that not all phonemes are equally difficult to

recognise for machines. More detailed analyses showed that, for instance, *voicing* information is difficult for machines to 'recognise', while *plosives* are relatively easy. This suggests that some important acoustic cues, used by human listeners, are not present in the acoustic features created by the front-end of the ASR system. Further research is needed to identify which cues are used by human listeners and are not present in the acoustic features. Subsequently, the acoustic features should be adapted such that they do contain those important acoustic cues.

How do infants learn speech? How do they learn to segment the speech? How do they learn what these segments should be? How do they learn new words? And how do they learn how to distinguish between non-speech background sounds and the actual speech? These are issues that are so easy for a child to resolve, but as yet there is no complete understanding (and thus no computational model) of this process. Since infants acquire language so incredibly quickly and well, it is important to keep exploring the potential benefit of using knowledge about language acquisition for the improvement of ASR systems and computational models of HSR (current computational models of human speech recognition are like ASR systems in that they also use a predefined lexicon with predefined sub-word units to represent the lexical items). How then can this knowledge about child language acquisition be used to improve ASR systems and computational models of HSR? It is well possible that once it is understood how infants acquire language this will lead to the need of totally new architectures for the automatic recognition of speech (even beyond the probabilistic pattern recognition techniques currently in use). For instance, when a child acquires language, the units into which the acoustic signal will be segmented are not pre-specified as is currently the case for ASR systems and computational models of HSR. This necessitates the development of a new architecture for ASR systems and computational models that makes use of *emergent* units of recognition – not necessarily in terms of the linguistically-based recognition units used in current ASR systems and computational models.

In order to provide answers to the above questions and issues, it is necessary that researchers from both fields start to collaborate (more); cognitively plausible computational models of human speech recognition such as CELL (Roy & Pentland, 2002) and SpeM (Scharenborg et al., 2005) provide an excellent and useful starting point. We have just started to get to know one another, now it is time to make things work. There is a bright future for research in-between the fields of human and automatic speech recognition.

7. Acknowledgements

This research was supported by a Talent stipend from the Netherlands Organization for Scientific Research (NWO). The author is greatly indebted to Louis ten Bosch and Lou Boves (Radboud University Nijmegen, The Netherlands), Sue Harding and Martin Cooke (University of Sheffield, UK), and two anonymous reviewers for their comments on earlier versions of this manuscript.

8. References

Allen, J.B., 2004. Consonant recognition and the articulation index. Journal of the Acoustical Society of America, 117 (4), 2212-2223.

Alsteris, L.D., Paliwal, K.K., 2006. Further intelligibility results from human listening tests using the short-time phase spectrum. Speech Communication, 48 (6), 727-736.

Axelrod, S., Mason, B., 2004. Combination of hidden Markov models with dynamic time warping for speech recognition. Proceedings of IEEE ICASSP, Montreal, Canada, pp. 173-176.

Barker, J.P., Cooke, M., Ellis, D.P.W., 2005. Decoding speech in the presence of other sources, Speech communication, 45, 5-25.

Bourlard, H., Hermansky, H., Morgan, N., 1996. Towards increasing speech recognition error rates. Speech Communication, 18, 205-231.

Bregman, A.S., 1990. Auditory scene analysis: The perceptual organization of sound. MIT Press, Cambridge, MA.

Carey, M.J., Quang, T.P., 2005. A speech similarity distance weighting for robust recognition. Proceedings of Interspeech, Lisbon, Portugal, pp. 1257-1260.

Carpenter, B., 1999. Human versus machine: Psycholinguistics meets ASR. Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, CO, pp. 225-228.

Chomsky, N., Halle, M., 1968. The sound pattern of English. MIT Press, Cambridge, MA.

Cole, R., Fanty, M., Muthusamy, Y., Gopalakrishnan, M., 1990. Speaker-independent recognition of spoken English letters. Proceedings IEEE INNS International Joint Conf. on Neural Networks, San Diego, CA, Vol. 2, pp. 45-51.

Cooke, M., 2006. A glimpsing model of speech recognition in noise. Journal of the Acoustical Society of America, 119 (3), 1562-1573.

Cooke, M., Ellis, D.P.W., 2001. The auditory organization of speech and other sources in listeners and computational models. Speech Communication, 35, 141-177.

Cooke, M., Green, P., Josifovski, L, Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication, 34, 267-285.

Cooke, M., Green, P.G., Crawford, M.D., 1994. Handling missing data in speech recognition. Proceedings of ICSLP, Yokohama, Japan, pp.1555-1558.

Cutler, A., Robinson, T., 1992. Response time as a metric for comparison of speech recognition by humans and machines. Proceedings of ICSLP, Banff, Canada, pp. 189-192.

Davis, M.H., Marslen-Wilson, W.D., Gaskell, M.G., 2002. Leading up the lexical garden-path: Segmentation and ambiguity in spoken word recognition. Journal of Experimental Psychology: Human Recognition and Performance, 28, 218-244.

Davis, S. Mermelstein, P., 1980. Comparison of the parametric representation for monosyllabic word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357-366.

de Boer, B, Kuhl, P.K., 2003. Investigating the role of infant-directed speech with a computer model, ARLO, 4, 129-134.

De Wachter, M., Demuynck, K., van Compernolle, D., Wambaq, P., 2003. Data driven example based continuous speech recognition. Proceedings of Eurospeech, Geneva, Switzerland, pp. 1133-1136.

Dusan, S., Rabiner, L.R., 2005. On integrating insights from human speech recognition into automatic speech recognition. Proceedings of Interspeech, Lisbon, Portugal, pp. 1233-1236.

Furui, S., 2001. From read speech recognition to spontaneous speech understanding. In R. Smits, J. Kingston, T.M. Nearey & R. Zondervan (Eds.), Proceedings of the workshop on speech recognition as pattern classification (pp. 55-60). Nijmegen, MPI for Psycholinguistics.

Gaskell, M.G., Marslen-Wilson, W.D., 1997. Integrating form and meaning: A distributed model of speech recognition. Language and Cognitive Processes, 12, 613-656.

Goldinger, S.D., 1998. Echoes of echoes?: An episodic theory of lexical access. Psychological Review, 105, 251-279.

Harley, T., 2001. The psychology of language. From data to theory. Psychology Press, Hove.

Hawkins, J., 2004. On Intelligence. Times Books, New York, NY.

Hawkins, S., 2003. Roles and representations of systematic fine phonetic detail in speech understanding. Journal of Phonetics, 31, 373-405.

Hermansky, H., 1998. Should recognizers have ears? Speech Communication, 25, 3-27.

Hermansky, H., 2001. Human speech recognition: Some lessons from automatic speech recognition. In R. Smits, J. Kingston, T.M. Nearey & R. Zondervan (Eds.), Proceedings of the workshop on speech recognition as pattern classification (pp. 61-66). Nijmegen, MPI for Psycholinguistics.

Hintzman, D.L., 1984. MINERVA2: A simulation model of human memory. Behavior Research Methods, Instruments, & Computers, 16, 96-101.

Holmes, J., Holmes, W., 2002. Speech recognition and synthesis. Taylor & Francis, London, UK.

Huckvale, M., 1998. Opportunities for re-convergence of engineering and cognitive science accounts of Spoken word recognition. Proceedings of the Institute of Acoustics Conference Speech and Hearing, Windermere, pp. 9-20.

Kemps, R.J.J.K., Ernestus, M., Schreuder, R., Baayen, R.H., 2005. Prosodic cues for morphological complexity: The case of Dutch plural nouns. Memory & Cognition, 33, 430-446.

King, S., Taylor, P., 2000. Detection of phonological features in continuous speech using neural networks. Computer Speech and Language, 14, 333-353.

Kirchhoff, K., 1999. Robust speech recognition using articulatory information, Ph.D. thesis, University of Bielefield.

Kirchhoff, K., Schimmel, S., 2005. Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. Journal of the Acoustical Society of America, 117 (4), 2238-2246.

Klatt, D.H., 1979. Speech recognition: A model of acoustic-phonetic analysis and lexical access. Journal of Phonetics, 7, 279-312.

Klatt, D.H., 1989. Review of selected models of speech recognition. In W.D. Marslen-Wilson (Ed.), Lexical representation and process (pp. 169-226). Cambridge, MA: MIT Press.

Krause, J.C., Braida, L.D., 2002. Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. Journal of the Acoustical Society of America, 112 (5), 2165-2172.

LDC, 1995. SWITCHBOARD: A User's Manual, Catalog Number LDC94S7. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Leonard, R.G., 1984. A database for speaker-independent digit recognition. Proceedings of IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 42.11.1-42.11.4.

Lippmann, R., 1997. Speech recognition by machines and humans. Speech Communication, 22 (1), 1-15.

Livescu, K., Glass, J., Bilmes, J., 2003. Hidden feature models for speech recognition using dynamic Bayesian networks. Proceedings of Eurospeech, Geneva, Switzerland, pp. 2529-2532.

Luce, P.A., Goldinger, S.D., Auer, E.T., Vitevitch, M.S., 2000. Phonetic priming, neighborhood activation, and PARSYN. Recognition & Psychophysics, 62, 615-625.

Maier, V., Moore, R.K., 2005. An investigation into a simulation of episodic memory for automatic speech recognition. Proceedings of Interspeech, Lisbon, Portugal, pp. 1245-1248.

Marslen-Wilson, W. D., 1987. Functional parallelism in spoken word recognition. Cognition, 25, 71-102.

McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech recognition. Cognitive Psychology, 18, 1-86.

McQueen, J.M., 2004. Speech recognition. In K. Lamberts & R. Goldstone (Eds.), The handbook of cognition (pp. 255-275). London: Sage Publications.

Meyer, B., Wesker, T., Brand, T., Mertins, A., Kollmeier, B., 2006. A human-machine comparison in speech recognition based on a logatome corpus. Proceedings of the workshop on Speech Recognition and Intrinsic Variation, Toulouse, France.

Miller, G.A., Nicely, P.E., 1954. An analysis of perceptual confusions among some English consonants. Journal of the Acoustical Society of America, 27 (2), 338-352.

Moore, R.K., 1995. Computational phonetics. Proceedings of XIIIth International Congress of Phonetic Sciences, Stockholm, Sweden.

Moore, R.K., 2001. There's no data like more data: but when will enough be enough? Proceedings of the Acoustics Workshop on Innovations in Speech Processing, 23 (3), Stratford-upon-Avon, UK, pp.19-26.

Moore, R.K., 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners. Proceedings of Eurospeech, Geneva, Switzerland, pp. 2581-2584.

Moore, R.K., Cunningham, S.P., 2005. Plasticity in systems for automatic speech recognition: a review. Proceedings of the ISCA Workshop on Plasticity in Speech Recognition, London, UK, pp. 109-112.

Moore, R.K., Cutler, A., 2001. Constraints on theories of human vs. machine recognition of speech. In R. Smits, J. Kingston, T.M. Nearey & R. Zondervan (Eds.), Proceedings of the workshop on speech recognition as pattern classification (pp. 145-150). Nijmegen, MPI for Psycholinguistics.

Nearey, T.M., 2001. Towards modelling the recognition of variable-length phonetic strings. In R. Smits, J. Kingston, T.M. Nearey & R. Zondervan (Eds.), Proceedings of the workshop on speech recognition as pattern classification (pp. 133-138). Nijmegen, MPI for Psycholinguistics.

Norris, D., 1994. Shortlist: A connectionist model of continuous speech recognition. Cognition, 52, 189-234.

Norris, D., 2005. How do computational models help us develop better theories? In A. Cutler (Ed.), Twenty-first century psycholinguistics: Four cornerstones (331-346). Hillsdale, NJ: Erlbaum.

Ostendorf, M., 1999. Moving beyond the 'beads-on-a-string' model of speech. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, Keystone, CO, pp. 79-84.

Paul, D., Baker, J., 1992. The design for the Wall Street Journal-based CSR corpus. Proceedings DARPA Speech and Natural Language Workshop. Morgan Kaufmann, Austin, TX, pp. 357-360.

Pols, L.C.W., 1999. Flexible, robust, and efficient human speech processing versus present-day speech technology. Proceedings of ICPhS, San Francisco, CA, pp. 9-16.

Qin, M.K., Oxenham, A.J., 2003. Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers, Journal of the Acoustical Society of America, 114 (1), 446-454.

Rabiner, L., Juang, B.-H., 1993. Fundamentals of speech recognition. New Jersey: Prentice Hall.

Raj, B., Seltzer, M., Stern, R., 2004. Reconstruction of missing features for robust speech recognition. Speech Communication, 43 (4), 275-296.

Roy, D.K., Pentland, A.P., 2002. Learning words from sights and sounds: A computational model. Cognitive Science, 26, 113-146.

Salverda, A.P., Dahan, D., McQueen, J.M., 2003. The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. Cognition, 90, 51-89.

Scharenborg, O., 2005a. Parallels between HSR and ASR: How ASR can contribute to HSR. Proceedings of Interspeech, Lisbon, Portugal, pp. 1237-1240.

Scharenborg, O., 2005b. Narrowing the gap between automatic and human word recognition. Ph.D. thesis, Radboud University Nijmegen, The Netherlands.

Scharenborg, O., Norris, D., ten Bosch, L., McQueen, J.M., 2005. How should a speech recognizer work? Cognitive Science, 29 (6), 867-918.

Scharenborg, O. ten Bosch L, Boves, L., 2007. 'Early recognition' of polysyllabic words in continuous speech. Computer Speech and Language, 21 (1), 54-71.

Scharenborg, O., ten Bosch, L., Boves, L., Norris, D., 2003. Bridging automatic speech recognition and psycholinguistics: Extending Shortlist to an end-to-end model of human speech recognition. Journal of the Acoustical Society of America, 114 (6), 3032-3035.

Scharenborg, O., Wan, V., Moore, R.K., 2006. Capturing fine-phonetic variation in speech through automatic classification of articulatory features. Proceedings of the workshop on Speech Recognition and Intrinsic Variation, Toulouse, France, pp. 77-82.

Seltzer, M., Raj, B., Stern, R., 2004. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. Speech Communication, 43, 379-393.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues, Science, 270, 303-304.

Srinivasan, S., Wang, D., 2005. Modeling the recognition of multitalker speech. Proceedings of Interspeech, Lisbon, Portugal, pp. 1265-1268.

Sroka, J.J., Braida, L.D., 2005. Human and machine consonant recognition. Speech Communication, 45, 401-423.

Strik, H., 2003. Speech is like a box of chocolates. Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain, pp. 227-230.

Strik, H., 2006. How to handle pronunciation variation in ASR: By storing episodes in memory? Proceedings of the workshop on Speech Recognition and Intrinsic Variation, Toulouse, France, pp. 33-38.

Tanaka, H., 2000. Turn-Projection in Japanese Talk-in-Interaction. Research on Language and Social Interaction, 33 (1), 1-38.

ten Bosch, L., 2001. ASR-HSR from an ASR point of view. In R. Smits, J. Kingston, T.M. Nearey & R. Zondervan (Eds.), Proceedings of the workshop on speech recognition as pattern classification (pp. 49-54). Nijmegen, MPI for Psycholinguistics.

Tuller, B., 2003. Computational models in speech recognition. Journal of Phonetics, 31, 503-507.

van Leeuwen, D.A., van den Berg, L.G., Steeneken, H.J.M., 1995. Human benchmarks for speaker independent large vocabulary recognition performance. Proceedings of Eurospeech, Madrid, Spain, pp. 1461-1464.

Voskuhl, A., 2004. Humans, machines, and conversations: An ethnographic study of the making of automatic speech recognition technologies. Social Studies of Science, 34 (3), 393-421.

Wade, T., Eakin, D.K., Webb, R., Agah, A., Brown, F., Jongman, A, Gauch, J., Schreiber, T.A., Sereno, J., 2002. Modeling recognition of speech sounds with Minerva2, Proceedings of ICSLP, Denver CO, pp. 1653-1656.

Wang, D.L., Brown, G.J. (Eds.), 2006. Computational auditory scene analysis: principles, algorithms, and applications. IEEE Press/Wiley-Interscience.

Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., Kollmeier, B., 2005. Oldenburg Logatome speech corpus (OLLO) for speech recognition experiments with humans and machines. Proceedings of Interspeech, Lisbon, Portugal, pp. 1273-1276.

Wester, M., 2003. Syllable classification using articulatory-acoustic features. Proceedings of Eurospeech, Geneva, Switzerland, pp. 233-236.

Wester, M., Greenberg, S., Chang, S., 2001. A Dutch treatment of an Elitist approach to articulatory-acoustic feature classification. Proceedings of Eurospeech, Aalborg, Denmark, pp. 1729-1732.

Wright, R., 2006. Intra-speaker variation and units in human speech perception and ASR. Proceedings of the workshop on Speech Recognition and Intrinsic Variation, Toulouse, France, pp. 39-42.

Yu, C., Ballard, D.H., Aslin, R.N., 2005. The role of embodied intention in early lexical acquisition, Cognitive Science, 29(6), 961-1005.