Parallels between HSR and ASR: How ASR can Contribute to HSR

Odette Scharenborg

CLST, Radboud University Nijmegen, The Netherlands
O.Scharenborg@let.ru.nl

Abstract

In this paper, we illustrate the close parallels between the research fields of human speech recognition (HSR) and automatic speech recognition (ASR) using a computational model of human word recognition, SpeM, which was built using techniques from ASR. We show that ASR has proven to be useful for improving models of HSR by relieving them of some of their shortcomings. However, in order to build an integrated computational model of all aspects of HSR, a lot of issues remain to be resolved. In this process, ASR algorithms and techniques definitely can play an important role.

1. Introduction

Two research fields that investigate (parts of) the speech recognition process are automatic speech recognition (ASR) and human speech recognition (HSR). Although the two research areas are closely related their aims and research approaches are different. In ASR, the central issue is minimising the number of recognition errors. Much research effort in ASR has therefore been put into the development of systems that generate accurate lexical transcriptions of acoustic speech signals. In HSR research, the goal is to understand how listeners recognise spoken utterances. This is done by creating theories and building computational models of HSR, which can be used for the simulation and explanation of the human speech recognition process. In this paper, the focus is on symbolic theories and models of HSR.

Although both ASR and HSR claim to investigate the whole recognition process from the acoustic signal to the recognised units, an ASR system necessarily is an end-to-end system – it must be able to recognise words from the acoustic signal – while most models of HSR describe only parts of the human speech recognition process. An integral model covering all stages of the human speech recognition process does not yet exist. One part of the recognition process that virtually all models of human speech recognition lack is the part that converts the acoustic signal into some kind of discrete symbolic representation. Consequently, most existing HSR models cannot recognise real speech. This makes it hard to evaluate the theoretical assumptions underlying models of HSR in real-life test conditions.

Despite the gap that seperates the two fields, there is a growing interest in possible cross-fertilisation (e.g., [1]). Some strands in HSR research hope to deploy ASR approaches to integrate partial modules into an end-to-end model [2]. From the point of view of ASR, there is some hope to improve performance by incorporating essential knowledge about HSR into current ASR systems ([3], [4]).

The aim of this paper is two-fold. The first aim is to reveal the close parallels between the fields of HSR and ASR with respect to the speech recognition process. The second aim of this paper is to illustrate in more detail how ASR can contribute to building a convincing end-to-end computational model of all aspects of the human speech recognition process.

We will illustrate the close parallels by comparing the implementations of current computational models of HSR and SpeM. SpeM is a computational model of human word recognition built using techniques from the field of ASR that is able to recognise real speech [5]. We will further illustrate the existence of the close parallels by explaining important issues that need to be dealt with when building an integrated model of HSR, and we will describe how this is done in SpeM. In the second part of this paper, we will describe several of the issues that remain to be solved in order to build an integrated model of all aspects of the human speech recognition process, and how algorithms and techniques known from ASR may contribute to solve these issues. This endeavour will further narrow the gap that has existed for decades between the research fields of HSR and ASR.

2. Revealing the close parallels

2.1. The prelexical level

Symbolic theories of HSR claim 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 (e.g., [6], [7], [8]). According to symbolic theories, the speech recognition process thus consists of two levels: the prelexical level and the lexical level. A central requirement of symbolic computational models is thus an intermediate segmental representation of the speech signal. However, as explained before, most HSR models lack a module that converts the speech signal into a segmental representation; instead they use a handcrafted ‘error-free’ discrete representation of the input – in the sense that the input always perfectly aligns with the segmental representations of the words in the lexicon. Thus in effect, in most symbolic computational models, the process of creating the prelexical representations is only assumed, and not physically present. The output of the prelexical process is available in the form of the handcrafted segmental representation of the speech signal.

This property could, however, be irrelevant if such an ‘error-free’ representation of the speech signal could be generated automatically. The handcrafted input could then be replaced by the ‘real’ representation of the speech signal. But is it at all likely that such an ‘error-free’ discrete representation of the speech signal can be (automatically) created? There are reasons to believe that a unique segmental representation of the speech signal does not exist. One of these reasons is that no absolute truth exists as to what phones a person has produced; therefore, it is not possible to obtain a unique and ‘true’ symbolic transcription of a given speech signal [9]. Furthermore, studies in phonetics “suggested that the more detailed a transcription is, the less reliable it tends to be” [10]. This statement is backed-up by experiments

described in, e.g., [11]. They report on a consensus transcription procedure. Two experienced transcribers created a narrow consensus transcription of continuous speech samples. Six weeks after the last tape had been transcribed they created a new narrow consensus transcriptions of 25 utterances for each of eight randomly selected speech samples. Four weeks later, another eight speech samples were randomly selected and transcribed. Comparing the original consensus transcriptions and the retest transcriptions segment by segment yielded an agreement of 68%. However, the percentage agreement went up to 76% when the diacritics were removed from the transcriptions. Therefore, it seems that the ideal segmental representation of the speech signal cannot be generated, and thus that the 'error-free' discrete segmental representation of the speech signal required by most models of HSR cannot be created on the basis of real speech. On top of that, HSR experiments (see [12], for an overview) have shown that the representations at the prelexical level should be probabilistic rather than categorical.

In short, when trying to build an integrated computational model of human speech recognition, the first two issues that need to be resolved are that the integrated model should contain a real module that simulates the prelexical level, and the output of the prelexical level should be probabilistic instead of categorical.

As a first step towards an integrated model of human speech recognition, SpeM (SPEech-based Model of human speech recognition, [5]) was developed. SpeM is an end-to-end model of human *word* recognition based on the theory underlying the Shortlist model [8], and was built using techniques from ASR. SpeM is not just a re-implementation of Shortlist; it represents an important advancement over existing models of HSR in that it is able to recognise real-life speech input at reasonably high levels of accuracy (see [5] for experimental results).

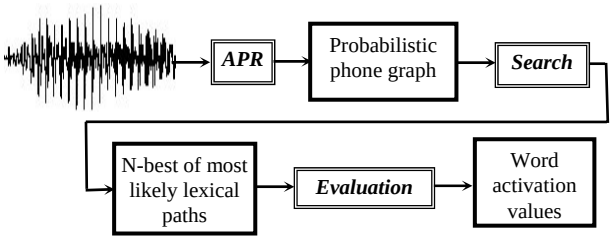

Figure 1. Graphical representation of the SpeM model.

Figure 1 shows a graphical representation of the SpeM model. SpeM consists of three modules that operate in sequence. The first module is an automatic phone recogniser (APR) which represents the prelexical level. The subsequent modules represent the lexical level. The input to the prelexical level is the acoustic signal. Using statistical acoustic models used in standard ASR systems, the acoustic signal is converted into a segmental representation of the speech signal (for more information on ASR systems and search, e.g., [13]). However, in contrast to the categorical linear representation of the speech signal used in most existing models of HSR, SpeM creates a probabilistic representation of the speech signal in the form of a probabilistic phone graph (Shortlist uses phonemes to represent the speech signal). The lexical search module searches for the word (sequence) that corresponds to the best path through the probabilistic phone lattice and a lexicon represented in the form of a lexical tree. The output is in the form of a list of the *N*-best paths through the phone lattice. The third module compares these alternative paths and computes a measure of the probability that, for a given input, individual words will be recognised. (The two modules associated with the lexical level will be discussed in more detail in the following section.)

In conclusion, the existence of the probabilistic prelexical processing in SpeM in the form of an APR built with ASR techniques solves the two issues (the need for the categorical input representation and the absence of a physically-present prelexical level) described in this section. It shows how HSR can benefit from techniques known from the field of ASR.

2.2. The lexical level

According to symbolic theories of HSR, the prelexical representations are mapped onto lexical representations by some kind of lexical processing. During the human speech recognition process, each incoming phoneme (or, alternatively, a set of phonetic features) is matched against the segmental representations of all words in an internal lexicon. By this process, all words that are roughly consistent with the bottom-up input are activated. The amount of activation of each word hypothesis is based on its degree of fit with the input. Finally, the word hypotheses that overlap in time in the input inhibit each other. This process is referred to as (lexical) competition. The activation of a word at a certain point in time is based on its initial activation and the inhibition caused by other activated words.

Data obtained in HSR experiments mostly 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 is able to provide a measure (usually referred to as '(word) activation') associated with the strength of different lexical hypotheses over time. The word activation score, then, can be compared to the performance of listeners in experiments where they are required to make word-based decisions. The word with the highest activation is ultimately recognised.

The search module of SpeM *computes* the bottom-up goodness-of-fit of different lexical hypotheses to the current input, while the evaluation module *acts to compare* those hypotheses with each other. During the search process, the best path (the optimal sequence of words) is derived using a time-synchronous Viterbi search through a search space. The search space is defined as the product of a lexicon (represented as a lexical tree) and the probabilistic phone graph. In a single forward pass, all nodes of the phone lattice are processed from left-to-right, and all hypotheses are considered in parallel. The words hypothesised by the search module are each assigned a score that corresponds to the degree of match of the word to the current input. Whenever the mismatch between the hypothesised word (and its history) and the input becomes too large, the hypothesis drops out of the *beam*, i.e., it is pruned away. As in ASR systems and similar to human speech recognition: only the most plausible paths are therefore considered. The output of the search module in SpeM is a ranked *N*-best list of alternative paths, each with an associated path score.

The evaluation module provides a procedure to derive a measure of *word* activation from the *path* scores calculated by SpeM. In SpeM, the word activation of a word *W* is closely related to the probability $P(W|X)$ of observing a word *W*, given the signal *X*. This can be rewritten using Bayes' Rule:

$$P(W \mid X) = \frac{P(X \mid W)P(W)}{P(X)}, \quad (1)$$

in which $P(W)$ is the prior probability of *W*, and $P(X)$ denotes the prior probability of observing the signal *X* (for details, see [14]). Bayes' Rule and the probability $P(W|X)$ play a central role in the mathematical framework on which statistical pattern matching techniques are built (i.e., most ASR implementations). The Bayesian decomposition of the probability $P(W|X)$ is the foundation on which we based the calculation of word activation.

The word activation as calculated by SpeM is not based on 'active' inhibition (like the inhibition between lexical representations in the Shortlist model). It models competition between words in a 'static' way. In [5], three simulations were run. The first simulation was based on research presented in [8] concerning the lexical embedding problem (the fact that any stretch of speech is likely to be consistent with several different lexical hypotheses) and the segmentation problem (how can continuous speech be segmented into words when there are no fully reliable cues to word boundaries?). The second and third simulations were concerned with lexical competition [15] and the Possible Word Constraint [16] (a parse that contains a (sequence of) phone(s) occurring between a target word and a boundary that is not phonotactically well formed, and thus not a possible word, is penalised), respectively. The results of these three simulations showed that SpeM was able to model correctly the outcomes of these three psycholinguistic studies.

The question remains, however, as to whether the current way of modelling competition suffices or whether an active inhibition is indeed necessary. It may be that the results of previous and certainly of future psycholinguistic studies can only be accounted for by assuming active inhibition. If that happens, the word activation calculation procedure in SpeM must be adapted and the issue that arises then is how this active inhibition should be implemented.

In conclusion, the successful implementation of a word activation score based on the path-based scores used in ASR search implementations, again shows the close ties between ASR and HSR. Additionally, almost all psychological models assume that human listeners can perform the search for words in parallel, but existing HSR models usually use a serial search. SpeM, however, is able to perform the search in parallel.

3. Towards an integrated model of HSR

The development of SpeM already showed the important contributions of ASR algorithms and techniques to the implementation of an end-to-end model of human word recognition. There are, however, more aspects to human speech recognition than those associated with word recognition. If one wants to build an integrated computational model of the human speech recognition process, all aspects of the human speech recognition process should be covered by one computational model. Since SpeM is able to recognise real speech (and is thus able to model a larger part of the human speech recognition process than most models of HSR), we take SpeM as the starting point of the next discussion about the development of an integrated model of all aspects of the human speech recognition process.

So far, we have only discussed research concerning word recognition. A different strand of HSR research is concerned with phoneme recognition. Experiments, in which listeners are required to make explicit phonemic judgements, show that lexical knowledge is used to make those judgements. For instance, phonemes are easier to spot in a real word than in a non-word (e.g., [17]). Secondly, an ambiguous phoneme stimulus on a word/non-word continuum is more likely to be classified in agreement with the word than the non-word [18]. For SpeM to be able to model the processes involved in phoneme recognition, and the lexical effects on phonetic perception, it has to be extended, for instance by adding a phoneme decision layer similar to the one implemented in Merge [19]. In Merge, both lexical and prelexical information can be used to make a judgement about the identity of a phoneme. In [14], SpeM was extended with a decision layer that was used to recognise a word before its acoustic offset. This decision layer could be adapted to account for the effects that are described here.

In [20], it is shown that the [raɪp] in 'right berries' where the /t/ is assimilated to a [p] is not identical to the [raɪp] in 'ripe berries'. Human listeners showed priming of the word 'right' but not of 'ripe' when the [raɪp] derived from 'right' was presented. Apparently, the assimilation process preserves usable acoustic-phonetic evidence about the unassimilated form of the word. These subphonemic cues appear to influence lexical activation. In the current version of SpeM, the subphonemic differences are implicitly available as log likelihood scores associated with the phones on the arcs of the phone graph generated by the APR. A lower log likelihood score means that the phone on the arc is less of a prototype of that phone class, perhaps due to assimilation processes. If one wants to model the influence of subphonemic cues on lexical activation explicitly, one possible solution would be to change the output of the APR into a graph of phonetic features (e.g., [21]) instead of phones.

As research by Goldinger [22] has shown, human listeners are able to remember details of specific tokens of words that they have heard. These memories for not only words but also speaker characteristics have shown to influence subsequent speech processing. One way for SpeM to be able to model the influence of these memories on the speech recognition process is to adapt the APR module such that it is able to use information about the speaker in building the phone graph and to pass speaker information to the lexical search module. A simple first step would be to train gender-dependent phone models for the APR, which can be used in parallel during the search. Gender-dependent phone models will improve the phone graph output by the APR. Furthermore, the output of the prelexical level then would contain information about the gender of the speaker. If a word has been spoken by a male, the acoustic models trained on male speech match the input better, resulting in a higher word activation for the target word. Secondly, the lexical influence of the memory of a word could be implemented using a 'dynamic' type of language model (LM), i.e., during speech recognition, for each previously recognised word the probability score in the LM could be increased. When the word is encountered for a subsequent time, it will receive a

higher word activation because of the higher probability of the word. If this method proves to work, it could be extended to specific groups of persons or individuals.

Humans are able to use contextual information in the speech recognition process. This contextual information is not just restricted to word frequency and/or the probability of co-occurrence of the current and the previous word (e.g., [23]). Experiments [24] have shown that context information is used after lexical access. For an integrated model of HSR to be able to simulate these results, LMs should be included. SpeM is able to use unigram and bigram LMs, which model the probability of co-occurrence of the current and the previous word. SpeM should thus be able to model effects found due to word frequency and the co-occurrence of two words (this, however, has not been tested yet). But, to be able to model the effects due to context information further away in the sentence (or even a previous sentence), SpeM should be extended such that it is able to use higher-order LMs. One could think of a strategy in which the recognition of a word boosts the probability of another word, as is found for humans.

The issues described in this section are not exhaustive, but they do illustrate the wide range of issues that remain to be solved in order to build an integrated model of all aspects of the human speech recognition process, and how algorithms and techniques from the field of ASR can contribute.

4. Concluding remarks

In this paper, several close parallels between the research fields of HSR and ASR were revealed. We showed that ASR has proven to be useful for improving models of HSR by relieving them of some of their shortcomings. We, therefore, believe that techniques and algorithms from the field of ASR can play an important role in order to build an integrated model of all aspects of the human speech recognition process.

5. Acknowledgements

The author would like to express her appreciation to Lou Boves, Louis ten Bosch, James M. McQueen, and Dennis Norris for fruitful discussions about the work presented in this paper, and the colleagues at CLST, in particular Lou Boves, for their comments on earlier versions of this paper.

6. References

[1] Moore, R.K., Cutler, A., "Constraints on theories of human vs. machine recognition of speech", *Proceedings of the SPRAAC workshop*, Nijmegen, The Netherlands, pp. 145-150, 2001.

[2] Nearey, T.M., "Towards modeling the perception of variable-length phonetic strings", *Proceedings of the SPRAAC workshop*, Nijmegen, The Netherlands, pp. 133-138, 2001.

[3] Carpenter, B., "Human versus machine: Psycholinguistics meets ASR", *Proceedings of ASRU*, Keystone, CO, pp. 225-228, 1999.

[4] Hermansky, H., "Human speech perception: Some lessons from automatic speech recognition", *Proceedings of the SPRAAC workshop*, Nijmegen, The Netherlands, pp. 61-66, 2001.

[5] Scharenborg, O., Norris, D., ten Bosch, L., McQueen, J.M., "How should a speech recognizer work?", *Accepted for publication in Cognitive Science.*

[6] Gaskell, M.G., Marslen-Wilson, W.D., "Integrating form and meaning: A distributed model of speech perception", *Language and Cognitive Processes*, *12*, 613-656, 1997.

[7] McClelland, J.L., Elman, J.L., "The TRACE model of speech perception", *Cognitive Psychology*, *18*, 1-86, 1986.

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

[9] Cucchiarini, C., "Phonetic transcription: A methodological and empirical study", *Ph.D. thesis*, University of Nijmegen, The Netherlands, 1993.

[10] Ball, M.J., Rahilly, J., "Transcribing disordered speech: The segmental and prosodic layers", *Clinical Linguistics & Phonetics*, *16*, *No. 5*, 329-344, 2002.

[11] Shriberg, L.D., Kwiatkowski, J., Hoffmann, K., "A procedure for phonetic transcription by consensus.", *J. of Speech and Hearing Research*, *27*, 456-465, 1984.

[12] McQueen, J.M., Dahan, D., Cutler, A., "Continuity and gradedness in speech processing", In A.S. Meyer & N.O. Schiller (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 39-78). Berlin: Mouton de Gruyter, 2003.

[13] Jelinek, F., "Statistical methods for speech recognition", Cambridge, MA: MIT Press, 1997.

[14] Scharenborg, O., ten Bosch, L., Boves, L., "'Early Recognition' of Words in Continuous Speech", *Proceedings of ASRU*, US Virgin Islands, 2003.

[15] McQueen, J. M., Norris, D., Cutler, A., "Competition in spoken word recognition: Spotting words in other words", *J. of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 621-638, 1994.

[16] Norris, D., McQueen, J. M., Cutler, A., Butterfield, S., "The possible-word constraint in the segmentation of continuous speech", *Cognitive Psychology*, *34*, 191-243, 1997.

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

[18] Ganong, W.F., "Phonetic categorization in auditory word perception", *J. of Experimental Psychology: Human Perception and Performance*, *6*, 111-125, 1980.

[19] Norris, D., McQueen, J. M., Cutler, A., "Merging information in speech recognition: Feedback is never necessary", *Behavioral and Brain Sciences*, *23*, 299-325, 2000.

[20] Gow, D.W., Jr., "Does English coronal place assimilation create lexical ambiguity?", *J. of Experimental Psychology: Human Perception and Performance*, *28*, 163-179, 2002.

[21] Wester, M., Greenberg, S., Chang, S., "A Dutch treatment of an elitist approach to articulatory-acoustic feature classification", *Proceedings of Eurospeech*, pp. 1729-1732, 2001.

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

[23] Marslen-Wilson, W. D., "Functional parallelism in spoken word recognition", *Cognition*, *25*, 71-102, 1987.

[24] Zwitserlood, P., "The locus of the effects of sentential-semantic context in spoken-word processing", *Cognition*, *32*, 25-64, 1989.