Bridging automatic speech recognition and psycholinguistics: Extending Shortlist to an end-to-end model of human speech recognition[a] (L)

Odette Scharenborg,[b] Louis ten Bosch, and Lou Boves
A^{2}RT, Department of Language and Speech, University of Nijmegen, The Netherlands

Dennis Norris
Medical Research Council Cognition and Brain Sciences Unit, Cambridge, United Kingdom

(Received 10 December 2002; accepted for publication 25 August 2003)

This letter evaluates potential benefits of combining human speech recognition (HSR) and automatic speech recognition by building a joint model of an automatic phone recognizer (APR) and a computational model of HSR, *viz.*, Shortlist [Norris, Cognition **52**, 189–234 (1994)]. Experiments based on ''real-life'' speech highlight critical limitations posed by some of the simplifying assumptions made in models of human speech recognition. These limitations could be overcome by avoiding hard phone decisions at the output side of the APR, and by using a match between the input and the internal lexicon that flexibly copes with deviations from canonical phonemic representations. © *2003 Acoustical Society of America.* [DOI: 10.1121/1.1624065]

PACS numbers: 43.71.An, 43.71.Es, 43.72.Ne [DOS] Pages: 3032–3035

I. INTRODUCTION

In this letter, we address speech recognition by making a bridge between two disciplines that have little overlap with respect to theoretical framework and experimental paradigms. One discipline is automatic speech recognition (ASR), which studies the automatic transformation of a speech signal into a sequence of discrete ''recognition tokens'' (commonly words). The main goal in ASR research is to minimize the number of recognition errors on a certain test set under specific testing conditions. The second discipline is the area of human speech recognition (HSR). In HSR, the conversion from an acoustic signal to (a string of) words is studied with a focus on understanding the psychological processes underlying human word recognition, e.g., the word perception process *per se*.

In HSR experiments, the usual stimuli are carefully spoken utterances recorded in noiseless environments. On the basis of theories of HSR, several computational models have been developed to simulate data from experiments on human speech perception. These models compute word activations as the input unfolds over time, where activation can be related to the speed and accuracy with which human listeners can recognize words. However, the existing computational models of HSR model only parts of the human speech recognition process. Typically, one of the missing parts is a module that converts the acoustic speech signal into a representation that forms an appropriate input for the models, which almost invariably assume some kind of *symbolic* representation of the speech signal.

Most experimental studies of HSR are based on read speech; however, in the last few years, the focus is shifting towards (more) spontaneous speech. Much more than read speech, spontaneous speech is affected by articulatory processes such as assimilation and reduction. Since listeners are sensitive to this type of subtle subphonemic information (e.g., Gow, 2002; see Cutler, 1998, for an overview), and to durational differences in the input (Davis *et al.*, 2002), HSR models are now challenged to address the question of how the speech signal is mapped onto lexical representations in more detail. This is an area where established techniques from ASR could be useful in informing future research. Nearey (2001) suggests combining dynamic pattern recognition techniques from ASR with HSR models in order to be able to use ''detailed phonetic models [...] as front ends for reasonable models of lexical access.'' Nearey doubts that existing HSR models ''will work as advertised when attached to real phonetic transduction systems.''

The present letter presents the results of experiments that put Nearey's conjecture to the test by attempting to make a bridge between the two research areas by studying a combined ASR–HSR model (henceforth referred to as ''joint model'') that can be regarded as an end-to-end model of human speech recognition. The input for the computational model of HSR is provided by an automatic phone recognizer (APR). This HSR model is tested with input consisting of extemporaneous, ''real-life'' speech.

II. THE JOINT MODEL

The proposed joint model is a first step in the development of an end-to-end model of HSR. From the available computational models for human word recognition, we have

[a]Earlier results and parts of the research presented in this article are published in the Proceedings of the ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology, Estes Park, Colorado, 2002, and in the proceedings of the 7th International Conference on Spoken Language Processing, Denver, Colorado, USA, 2002.

[b]Electronic mail: o.scharenborg@let.kun.nl

chosen Shortlist (Norris, 1994) to use in the joint model, because it has been successfully applied to a wide range of data from studies of HSR.

The joint model works as follows. The APR decodes a speech signal into a sequence of phone symbols; Shortlist takes this sequence as input and generates a sorted word list. These processes are discussed in more detail below.

A. Automatic phone recognizer (APR)

For the APR, we trained 36 context-independent (hidden Markov) phone models, one silence model, one model for hesitations such as ''uh,'' and one noise model (Scharenborg *et al.*, 2002a). The APR decoding is based on a phone loop with optional silence preceding and following each phone, and is guided by a phone bigram. The APR output is a purely phonemic representation of the acoustic signal—without word boundaries.

B. Shortlist

In its present implementation, Shortlist itself is a two-stage model. In the first stage, the input (i.e., a sequence of phone symbols) is processed from left to right and an exhaustive search of the internal lexicon yields a shortlist of word candidates (max. 30 per phone position[1]) that roughly match the phonemic input processed so far. The ''activation'' of the words in the short list is determined by the ''degree of fit'' between the phones in the input and the string of phones specified in Shortlist's internal lexicon. For each phone in the input that matches the lexicon representation of a word, the word's activation is increased by 1; otherwise, the activation is reduced by the mismatch parameter (default value is 3). In the second stage—the competition stage—the candidates in the shortlist enter into a network where time-overlapping candidates compete with each other. The output consists of (a sequence of) the most activated word(s).

III. MATERIAL

A. Acoustic data

For training the APR, data from a Dutch telephone corpus (the Dutch Directory Assistance Corpus, DDAC) were used (Sturm *et al.*, 2000). DDAC contains telephone calls to the Dutch 118 Directory Assistance service. Most utterances consist of either one Dutch city name or ''ik weet het niet'' (''I don't know'') pronounced in isolation. Others may also contain disfluencies and longer connected speech fragments. From this corpus, an independent test set (DDAC-test) of 10 510 utterances comprising 11 523 words was selected.

B. Lexicons

The *baseline* lexicon of Shortlist consists of 2392 city names and ''ik weet het niet'' (''I don't know''). For each word in the lexicon, one unique ''canonical'' phonemic representation was available.

The psycholinguistic theory underlying Shortlist makes no claim about the manner in which humans cope with pronunciation variation. Specifically, there is nothing in the theory that promotes the exclusive use of citation forms in the mental lexicon. Therefore, in order to deal with pronunciation variation, we created a second lexicon (''*PronVar*'') with on average 2.6 pronunciation variants per word (Scharenborg *et al.*, 2002a).

IV. EXPERIMENT I: BASELINE

We investigated the performance of the joint model in a baseline experiment using the baseline lexicon. The input for Shortlist consists of the speech utterances of DDAC-test transcribed by the APR. The parameter settings of Shortlist are identical to those used in Norris (1994). The ''performance'' of the joint model was tested in terms of the ASR benchmarking method of recognition errors, rather than on the psycholinguistic benchmark of similarity to human performance. Thus, the performance measure in this study is word accuracy: the percentage of utterances for which the reference words (in DDAC-test) receive the highest activation value at the output of Shortlist.

With an accuracy of 23.5%, the performance of the joint model in this baseline experiment appears to be quite poor. Since the performance of Shortlist on *canonical* phone representations is close to 100%, this result shows that recognizing real-life speech is more difficult than recognizing ''perfect'' phonemic transcriptions. An error analysis reveals that the model has great difficulty in dealing with reduced forms: the APR output mostly comprises fewer (and sometimes also different) phones than the canonical representation stored in Shortlist's lexicon.

Two follow-up experiments were carried out. The aim of the experiments was to study the possible improvement of the joint model's baseline performance using two strategies: using a lexicon that accounts for pronunciation variation (experiment II), and adjusting the value of the mismatch parameter in Shortlist (experiment III).

V. EXPERIMENT II: ACCOUNTING FOR PRONUNCIATION VARIATION

The second experiment is identical to experiment I, except that the PronVar lexicon (including pronunciation variations) was used. Using PronVar, Shortlist's performance as a speech recognizer—reported in terms of word accuracy—increases substantially with 16.2% absolute to 39.7%. An error analysis reveals that there are few cases where the correct word is in the shortlist, but that a competitor receives a higher final activation. This finding suggests that, in the case of noncanonical input, the selection of correct lexical candidates into the shortlist is problematic. This problem is addressed in experiment III.

VI. EXPERIMENT III: ADJUSTING THE MISMATCH PARAMETER

Listeners are highly sensitive to any mismatch between input phones and the phonological representations of words; a mismatch of a single phonological feature can eliminate all signs that a word has been activated (e.g., McQueen *et al.*, 1999). Because of these findings, Shortlist weights mismatching information much more heavily than matching information. However, a high value of the mismatch parameter

TABLE I. Effect of $M=3.0$ and $M=0.0$ measured in terms of the accuracy and the percentage of utterances for which the correct word was present in the shortlist (% In shortlist). Two lexicons are used, *viz.* baseline and PronVar.

	Baseline lexicon		*PronVar lexicon*	
Mismatch	Accuracy (%)	In shortlist (%)	Accuracy (%)	In shortlist (%)
3.0	23.5	24.3	39.7	42.3
0.0	32.5	59.5	54.1	76.5

could actually impair recognition of real-life speech considerably, as even quite small deviations from the expected lexical representation might make a word unrecognizable.

In experiment III, we investigated the effect of ''canceling'' the mismatch penalty (M) by setting $M=0.0$ in a test with both lexicons (for a complete account of the experiment, see Scharenborg *et al.*, 2002b). Table I shows the results in terms of the percentage of utterances for which the correct word is present in the shortlist ('In shortlist'). In addition, we report the word accuracy of the joint model on the word recognition task.

The first row of Table I shows the results of experiment II for reference. As can be seen in Table I, using $M=0.0$ increases the model's performance with both lexicons compared to the default value $M=3.0$.

VII. GENERAL DISCUSSION

The aim of the research described in this letter is to build and evaluate an end-to-end computational model of HSR—based on a joint model of an APR and Shortlist—that takes acoustic recordings of real-life speech as input. Real-life speech is characterized by pronunciation variation, which leads to noncanonical phonemic representations. In order to study the effects of noncanonical input to Shortlist, we carried out three experiments. Experiment I was the baseline experiment. In short, experiment II showed that including pronunciation variants in the internal lexicon of Shortlist improves the ability of the joint model to deal with real-life input. Experiment III showed that the combination of a mismatch parameter value of 0.0 and the use of the lexicon containing pronunciation variants is best able to deal with the reduced phonemic forms encountered in real-life speech. This combination yields a recognition accuracy of 54.1%, which is more than twice the baseline performance.

The experiments show that a straightforward combination of an APR and Shortlist does not yield an end-to-end model of HSR that can deal satisfactorily with real-life input, despite the fact that the APR and Shortlist each perform well in their own domains. Apparently, one cannot take for granted that a combination of the best models of two sides yields the best overall end-to-end model. Perhaps this is not too surprising, since neither system was designed with the intention of being interfaced with the other. Nevertheless, these experiments illustrate the consequences of some of the simplifying assumptions made in Shortlist and other HSR models, and show the extent to which these assumptions need to be revised to produce genuine end-to-end models that will be able to deal with the pronunciation variation present in spontaneous speech.

One shortcoming of the joint model is that it makes ''hard'' decisions both at the level of input phones, and in the goodness-of-fit metric used in the search process. Shortlist requires a single string of phone symbols as input. This implies that the APR is forced to make hard decisions about the segmental representation of the speech signal based only on the acoustic information. Also for HSR (e.g., Gaskell *et al.*, 1998; McQueen *et al.*, 1999), data from experiments indicate that human listeners do not make hard decisions prior to lexical selection. This problem with Shortlist has been addressed in the Merge model (Norris *et al.*, 2000), which is derived from Shortlist. However, the present implementation of Merge can handle only very small lexicons. One can eliminate hard decisions in the input by representing the speech signal as a segment-based lattice containing multiple segment-string hypotheses. The subsequent word search or activation algorithm should make the final decision which phones were present by reranking the activated words or taking the first best.

The second level of hard decisions involves the word search process in Shortlist. This search matches input phone strings to the phone strings stored in the lexicon in a way that it is intolerant of deviations from the canonical form of words. This is exactly the problem highlighted by Nearey (2001) and is certainly an area where more flexible pattern-matching techniques (such as dynamic programming as commonly used in ASR) could play an important role in refining computational HSR models. Of course, the resulting refined model should still be able to simulate actual data of HSR experiments.

An important question to be borne in mind when assessing the results of our experiments is whether our conclusion would have been radically different had we been able to drive Shortlist with the output of a human ''phone recognizer'' rather than the APR or with the output of an APR optimized on the task. Cucchiarini *et al.* (2001) showed that automatically generated transcriptions of read speech are very similar to manual phonetic transcriptions created by expert phoneticians. Such transcriptions are to a large extent also noncanonical. Thus, transcriptions created by human expert transcribers would cause similar problems for HSR models. In Scharenborg *et al.* (2002b), it is shown that optimizing the APR settings in order to improve the balance between generating an input phone sequence that is close to the signal and at the same time meets the input criteria of Shortlist, does not improve the performance of the joint model. So, while our experiments may not provide a precise quantitative measure of the extent of the problems faced by Shortlist, the problems are real nonetheless.

Finally, we would like to raise an additional point.[2] A human being is able to identify a nonlexical token as a nonword. However, the joint model is not able to classify any input as a nonword, since it simply activates the nearest known word. Identification of a nonword could be made possible by using an activation *threshold*: when no lexical token exceeds the threshold, the system identifies a nonword. This is one topic for further research.

VIII. CONCLUSION

This letter describes a coupling of an automatic phone recognizer and a computational model of human word recognition, *viz.* Shortlist. The coupling helped to identify aspects of the two components of the joint model that need to be improved in order to build a comprehensive end-to-end computational model of HSR that is able to deal with real-life speech. One of the future research directions is extending the representation of the speech signal from a single linear input phone string to a probabilistic phone graph. This allows, in a natural way, the postponement of a hard decision to a point later in the word search process, which we believe is desirable. A second possibility of improvement lies in changing the current word search in Shortlist into a search algorithm based on dynamic programming techniques. By doing so, deviations from the canonical representations can be dealt with in a natural way.

ACKNOWLEDGMENTS

The authors would like to thank Anne Cutler, James McQueen, and Roel Smits for fruitful discussions about this research and their comments on earlier versions of this letter. Furthermore, the authors would like to thank the four anonymous reviewers for raising additional interesting issues and giving their useful comments on an earlier version of this letter.

[1]The number 30 is arbitrarily chosen; the exact value does not have a large effect on the performance of the model (Norris, 1994).

[2]This issue was raised by one of the anonymous reviewers of an earlier version of this letter.

Cucchiarini, C., Binnenpoorte, D. M., and Goddijn, S. M. A. (**2001**). ''Phonetic Transcriptions in the Spoken Dutch Corpus: How to combine efficiency and good transcription quality,'' Proceedings of Eurospeech, pp. 1679–1682.

Cutler, A. (**1998**). ''The Recognition of Spoken Words with Variable Representations,'' Proceedings of the ESCA Workshop on Sound Patterns of Spontaneous Speech, Aix-en-Provence, pp. 83–92.

Davis, M. H., Marslen-Wilson, W. D., and Gaskell, M. G. (**2002**). ''Leading up the lexical gardenpath: Segmentation and ambiguity in spoken word recognition,'' J. Exp. Psychol. Hum. Percept. Perform. **28**, 218–244.

Gaskell, M. G., and Marslen-Wilson, W. D. (**1998**). ''Mechanisms of phonological inference in speech perception,'' J. Exp. Psychol. Hum. Percept. Perform. **24**, 380–396.

Gow, Jr., D. W. (**2002**). ''Does English coronal place assimilation create lexical ambiguity,'' J. Exp. Psychol. Hum. Percept. Perform. **28**, 163–179.

McQueen, J. M., Norris, D., and Cutler, A. (**1999**). ''Lexical influence in phonetic decision-making: Evidence from subcategorical mismatches,'' J. Exp. Psychol. Hum. Percept. Perform. **25**, 1363–1389.

Nearey, T. M. (**2001**). ''Towards modelling the perception of variable-length phonetic strings,'' Proceedings of the SPRAAC Workshop, Nijmegen, pp. 133–138.

Norris, D. (**1994**). ''Shortlist: A connectionist model of continuous speech recognition,'' Cognition **52**, 189–234.

Norris, D., McQueen, J. M., and Cutler, A. (**2000**). ''Merging information in speech recognition: Feedback is never necessary,'' Behav. Brain Sci. **23**, 299–370.

Scharenborg, O., and Boves, L. (**2002a**). ''Pronunciation Variation Modelling in a Model of Human Word Recognition,'' in Proceedings of Workshop on Pronunciation Modeling and Lexicon Adaptation, Estes Park CO, pp. 65–70.

Scharenborg, O., Boves L., and de Veth, J. (**2002b**). ''ASR in a Human Word Recognition Model: Generating Phonemic Input for Shortlist,'' Proceedings of ICSLP, pp. 633–636.

Sturm, J., Kamperman, H., Boves, L., and den Os, E. (**2000**). ''Impact of speaking style and speaking task on acoustic models,'' Proceedings of ICSLP, pp. 361–364.