INTERSPEECH 2009 BRIGHTON

The Case for Case-Based Automatic Speech Recognition

Viktoria Maier, Roger K. Moore

Department of Speech and Hearing, University of Sheffield, Sheffield, United Kingdom
V.Maier@dcs.shef.ac.uk, r.k.moore@dcs.shef.ac.uk

Abstract

In order to avoid global parameter settings which are locally suboptimal, this paper argues for the inclusion of more knowledge (in particular procedural knowledge) into automatic speech recognition (ASR) systems. Two related fields provide inspiration for this new perspective: (a) 'cognitive architectures' indicate how experience with related problems can give rise to more (expert) knowledge, and (b) 'case-based reasoning' provides an extended framework which is relevant to any similarity-based recognition systems. The outcome of this analysis is a proposal for a new approach termed 'Case-Based ASR'.

Index Terms: speech modelling, case based reasoning, exemplar-based systems

10.21437/Interspeech.2009-560

1. Introduction

It has become apparent that the performance of state-of-the-art automatic speech recognition (ASR) systems is approaching an asymptote at a level that falls well short of that which is desirable for many advanced applications [1] let alone being comparable with the capabilities of a human listener [2]. As a consequence, a number of researchers are exploring the field of human speech recognition (HSR) in order to both better understand the nature of speech and to investigate the possibility that a simulation of the human speech recognition system might lead to more competitive and robust ASR [3].

One of the key research areas to emerge from this link between ASR and HSR is the instance/exemplar-based approach [5][6][7]. The interest in such systems stems from the fact that these detail-retaining systems are able to exploit fine-acoustic and -phonetic detail [8]. However, some important detailed information is missing even in exemplar-based systems, such as the knowledge which information is the most salient for a particular comparison.

It is argued in this paper that answers to these questions may lie outside of mainstream ASR and/or HSR research and, in particular, may be found in more general research fields such as 'cognitive architectures' and 'case-based reasoning'. By opening ASR towards these two research fields it is shown how exemplar-based models that exploit increasing amounts of detail can be expanded to include model parameters and settings currently used in any contemporary ASR model.

2. Knowledge in contemporary ASR systems

Learning in ASR models is highly restricted. For example, for hidden Markov models (HMMs), the 'knowledge' that is learned is the mean, variance and mixture weight of each of the Gaussians, as well as the state transition probabilities. Every other type of knowledge is not learned by the system, but set globally by the designer of the system. Such parameters for an HMM/GMM include for example number of states and number of Gaussians per state, amongst others. In an example-based system such empirically set knowledge is also manifold. For example in the 'temporal episodic memory model' (TEMM) [7] such parameters are the power factor (or kernel width) and the optimal feature normalisation techniques.

Such knowledge is usually set *globally* to optimise performance over a particular evaluation set, as it is known that optimal settings depend on the *precise task*, usually seen as analogous to the database used. However, the true implication is that such settings are dependent on precise (local) structures of the problem, and thus setting these parameters globally implies that they may be set suboptimally locally. This conclusion is reminiscent of the observation that is a basis of the new-found interest of some of the ASR [5][6][7] and HSR [9][10][11] researchers in 'exemplar based' approaches. The main reason for this rise in interest is that the flexibility and robustness exhibited by HSR is not able to be modelled adequately with an architecture based on pre-abstracted representations. An exemplar-based approach offers a mechanism for retaining and accessing the 'fine phonetic detail' [8] that is discarded in purely abstract representations such as hidden Markov models (HMMs): in order to maximise classification performance, experience with a related (i.e. similar) classification problem is of the essence.

The similarity is a factor which sets the system's centre of attention. Attention mechanisms are very influential in HSR; attention however is not static, but dynamic. Attention mechanisms are influenced by knowledge of where the most relevant (or salient) information for a task lies [12][13][14][15], and may lead to categorical perception [16].

In cognitive psychology, information about how to perform a particular task is referred to as 'procedural knowledge'. Procedural knowledge (i.e. the knowing *how*) is generally distinguished from 'declarative knowledge' (i.e. the knowing *that*[1]). Episodic, as well as semantic knowledge are seen as types of declarative knowledge. In a living system, procedural knowledge is acquired (internalised) through interaction with the environment; it is not an external setting. All contemporary ASR approaches lack such detailed procedural knowledge.

3. Knowledge in cognitive architectures

The field of 'cognitive architectures' (CAs) addresses the creation and understanding of synthetic agents that support the same capabilities as human beings, i.e. the underlying infrastructure of intelligent systems. CAs are very general, covering aspects that are constant over time and across different application domains, thereby unifying findings across a range of different research fields.

Of particular interest here, are the knowledge sources generally identified for such CAs [17]. Of these, some can be readily associated with knowledge types used in ASR. Table 1 lists the knowledge sources mentioned in [17] and, where applicable, identifies their ASR counterparts.

[1] For example knowing *that* Berlin is the capital of Germany

Table 1. *Knowledge sources in CAs (left column) and their counterparts in ASR (right column).*

Knowledge types in cognitive architectures …	… and their counterparts in ASR
knowledge from past (through remembering)	retainment of training data, even if in generalised form (e.g. after training)
knowledge from past (through learning)	optimization of parameters
knowledge about the environment (learned via perception)	test input, once associated with meaning (e.g. after classification, or once related to stored data)
knowledge of the implications of the current situation (gained from planning, reasoning and prediction)	e.g. language model
knowledge can be learned via communication	-

There are a number of observations that can be made from Table 1. First, almost all knowledge sources listed in CAs are also represented in ASR systems. Second, while it is possible to argue for the existence of counterparts in ASR, they are highly limited in nature and very simple. However, the fact that almost all knowledge sources identified in CAs can be said to be approximated in ASR systems offer an indication that, in order to build more robust and well performing ASR systems, it is necessary to increase the 'intelligence' associated with each knowledge source - and one way to do this is through the use of 'case-based reasoning'.

4. Case-based reasoning

4.1. Background

Case-based reasoning (CBR) [18][19] is a research field in Artificial Intelligence that is related to expert-based systems. CBR solves new problems by adapting previously successful solutions to similar problems, and is hence seen to link strongly with the process of abstraction in human beings. CBR has been shown to be a part of human problem solving [20][21] and is thus seen as an AI technique that is founded in human cognition.

In essence, a 'case' in CBR denotes a problem situation, which has been learned and solved in such a manner that it can help solve future problems.

The CBR process consists of a four-stage cycle:

- Retrieve the most similar cases
- Reuse the cases to attempt to solve the problem
- Revise the proposed solution, if necessary
- Retain a new solution as a part of a new case

In practice, the revision in current CBR is usually done by a human interacting with the system rather than automatically.

Learning is a natural by-product of problem solving in CBR systems [19] and, naturally, CBR systems favour learning from past experiences. However, as noted by Aamodt & Plaza [19] "*effective learning in CBR requires a well worked out set of methods in order to extract relevant knowledge from the experience*".

4.2. Relevance of CBR to ASR

In theory, CBR is applicable to any problem solving, including speech recognition. Much of CBR research addresses solutions to very specific applications; for speech, however, such specific knowledge has been studied in great detail in the field of ASR, but not in the field of CBR. Additionally, CBR, with its core principle of generalising reuse based on similarity, has strong links with minimum-distance approaches to classification. All such systems assess the similarity of the current problem to stored examples, and these known examples are effectively *reused* to find the relevant answer to the current problem. By use of a similarity comparison, CBR (just as minimum-distance classifiers) sets a 'centre of attention'[1] (COA). The COA is therefore a fundamental, shared property between CBR and exemplar-based minimum-distance systems. The difference is that CBR may retain more information, including procedural knowledge, on which to base its decision of COA. As a result, CBR's mechanism to find a concrete solution can depend on varied, distinct processes. Thus, experience how to best address a particular problem can be incorporated. This allows for an optimal use of available data for a particular problem.

It is believed that ASR would benefit from such additional knowledge as well. What CBR can lend to ASR is its general insight into the supporting framework which would allow the reuse of multiple sources of knowledge, such as procedural knowledge. Such a proposed framework for ASR is termed 'case-based ASR'.

5. Instance-based minimum-distance classifier: natural correlation with CBR

As already mentioned, instance-based minimum-distance classifiers are can be said to naturally include a rudimentary concept of attention, referred to as COA. In effect, the COA is set in two dimensions (Figure 1). Dimension one addresses which traces in memory are those that are the most relevant for classifying the current input (where traces denote a stored unit of speech, for example a frame). This is referred to as *vertical* COA. The second dimension addresses which part(s) of the speech signal (i.e. which features) is (are) the most relevant for classifying the current input. This will be referred to as *horizontal* COA.

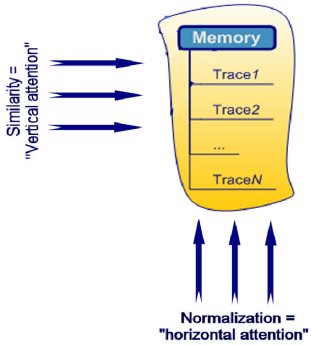

Figure 1: *Innate attention mechanisms in an instance-based minimum-distance classifier. A trace stands for a stored unit of speech.*

[1] COA is used here to refer to a specialised application of some of the known behaviours of human attention mechanisms for ASR, and is not intended to reflect any of the questions addressed in human attention research.

COA in such systems is defined globally by the chosen similarity (or activation) function as well as by the process of normalisation. Vertical COA is defined via the similarity function, and horizontal COA is set via 'normalised weighting'. Normalised weighting (different to normalisation) does not mean *equal* importance of features, but instead means the importance of features based on their salience for the particular speech task at hand. This means that instead of normalising features, an ASR system should focus on (i.e. pay attention to) features that are important for correct classification. Conventional normalisation will lead to suboptimal use of the relevant information.

5.1. Experience: from language-universal to language expert

In order to use the supplied information (i.e. features) to the fullest, experience is necessary. When an infant is born, its perception is language universal; it can discriminate equally well the phonetic details of any language [22]. Through linguistic *experience*, a child's perception is altered over time, and they become language experts [22]. Language experts know which parts of the sound input are those that are the most relevant, where attention should be centred on, resulting in categorical perception [16]. In this context, linguistic experience can be seen to reflect not only the fact that the infant has been exposed to language, but also that it has acquired *knowledge* and *skill* in dealing with a particular language.

Clearly, in order for an ASR system to achieve similar performance as a human listener, it needs to become a 'language-expert' system. This means that such a system must acquire and store experience that symbolises *expert knowledge* (which is largely procedural).

5.2. Knowledge in the new framework

A framework that incorporates knowledge along the lines mentioned above should be able to address many of the weaknesses of current ASR systems. In particular, such additional knowledge would address:

- horizontal attention (i.e. knowledge of feature importance), which should improve discrimination of most likely competitors in the system by focussing on those features most salient to the distinction between these classes
- the best parameter settings (e.g. power factor or kernel width) locally, based on previous experience (i.e. learning)
- the choice of similarity function locally, based on previous experience

Further weaknesses of current ASR systems which could be addressed with such a framework include:

- Strengthening the top-down influence by extending its influence on the vertical attention process
- New knowledge can be incorporated:
  - Knowledge of the world, including
  - Hierarchies of knowledge

Such knowledge, in particular procedural knowledge, brings such a system closer to the performance of a language expert, naturally including a more sophisticated form of attention mechanism.

6. Case-based ASR

It is argued that ASR has simplified important knowledge sources to a point where the knowledge retained can no longer address a particular (local) problem optimally. In order to create systems that can perform speech recognition as well as humans, more (expert) knowledge must be retained. For example, in a minimum-distance classifier how is each trace best separable from very similar neighbouring members of another class? In order to address this question in the ASR system an analysis of following information is necessary:

Vertical COA:

- Which other traces are very similar to the trace in question? Which are their classes?
- Which classes are the most relevant competitors?
- What is the prior probability of this trace being of a certain class?

Horizontal COA:

- Where is the most relevant difference in features between the most relevant traces of the various most likely classes?
- Which context information may be used best to distinguish highly likely competitors?

These questions help assess the requirements of a case-based framework for ASR. In an ASR database, examples are given as well as the word classes each contains. In order to supply additional sources of knowledge, the system needs to perform a deeper analysis of the available data, in particular to answer the points above.

Such an analysis can be performed in a more or less supervised fashion. One way would be that the system has a robust learning mechanism and can derive a high level of insight: all further information is learned by the system without external 'experience sharing'.

Another, simpler form of learning, would be to give the system a) the knowledge how to find optimal local parameters for particular functions and b) a choice of alternative functions to optimise the use of data in a particular problem. Here, the system only needs to analyse (offline, in the learning phase) which are its optimal system settings for a local problem. The goal of the system during such a learning phase is to maximise certain system criteria. Such system criteria could be i) finding the correct class and ii) increase the confidence of the best- and decreasing the confidence for the second-best class.

Such analyses of the data provided to the system via the database could be triggered 1) as a batch process when the database is first integrated into the ASR system; and/or 2) as corrective training when the system learns of a wrong assessment of an input, or when the system's certainty of a solution is not confident enough The difference to current corrective training procedures is that more/new knowledge types are corrected. Corrective training, different to the batch-training in 1) addresses a particular problem situation, concentrating on maximising the criteria for that particular case, until the understanding is such that the criterion that triggered the analysis is fulfilled.

This newly learned (expert) knowledge is then stored for reuse, in CBR fashion (i.e. based on the association of new knowledge types with particular COA's in the bottom-up data). This framework implements the four *CBR steps* (added in brackets) as follows:

- (CBR: *learning*, possibly *revise*) a learning algorithm to find optimal local handling of data
- Storage of such extended knowledge that does not fit the criteria to be stored in (CBR: *retain*) the traces, for example procedural knowledge
- A mechanism for (CBR: *retrieve*) access and (CBR: *reuse*) use of such extended knowledge.

The resulting framework suggestion is shown in Figure 2. Note that the graph does not include the learning step.

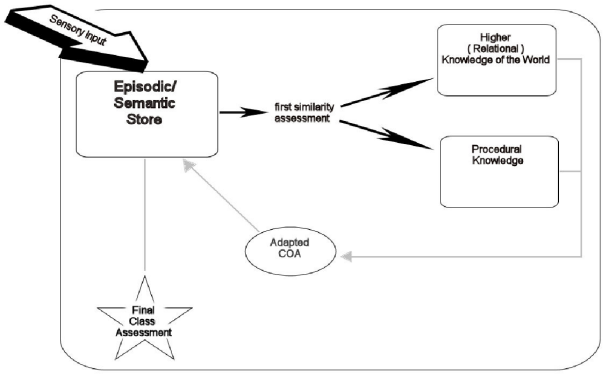

Figure 2: *Proposed framework: an input triggers a first analysis of the similarity to the known traces. The found activations of the traces activate the relevant information about a) the higher knowledge of the world and b) procedural knowledge. Stored experience can then help improve assessment of the data in the current step and/or future steps.*

7. Conclusions

This paper puts forward the position that current ASR systems handle many knowledge sources in a too simplified way by setting them to one global value. This leads to suboptimal handling of local problems. It is argued that in order to maximise performance locally, parameters should be set locally. Such maximised handling of a local problem is inspired by human processing. Humans possess not only declarative knowledge, but also procedural knowledge, and such knowledge needs to be acquired and stored in order to maximise local performance. Instead of setting parameters globally, such as the number of Gaussians per states or the number of states in an HMM, such optimal local settings should be learned by the system. One simple type of such a learning algorithm would be the learning of optimal parameters from alternatives in order to maximise system performance goals.

The need to enrich available knowledge sources, in particular procedural knowledge, in an ASR system is seen as generally applicable to the field of ASR. As such knowledge should be applied locally, it lends itself to combine multiple types of knowledge in a system via 'cases'. Such a system is thus referred to as case-based ASR. Cases are particularly incorporated in instance-based minimum-distance systems, which are highly related to CBR, with their core of exploitation of detail and setting of a COA via the distance metric.

8. Acknowledgement

This research was funded by the European Commission, under contract number FP6-034362, in the ACORNS project (www.acorns-project.org).

9. References

[1] Lippmann, R., "Speech Recognition by Machines and Humans", J. Speech Communication, 22, 1-15, Elsevier, 1997.

[2] Moore R. K., "A comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners", Proc. Eurospeech, 2582-2584, 2003.

[3] Moore, R. K. and Cutler, A., "Constraints on Theories of Human vs. Machine Recognition of Speech", Proc. SPRAAC Workshop on Human Speech Recognition as Pattern Classification, Max-Planck-Institute for Psycholinguistics, Nijmegen, 2001.

[4] Hintzman, D. L., "Schema-Abstraction in a Multiple-Trace Memory Model", Psychological Review, 93: 411-427, 1986.

[5] De Wachter, M. "Example Based Continuous Speech Recognition". PhD thesis, K.U.Leuven, ESAT, May 2007.

[6] Maier, V., Moore, R. K., "An Investigation into a Simulation of Episodic Memory for Automatic Speech Recognition", Proc. Interspeech, 2005.

[7] Maier, V., Moore, R. K., "Temporal Episodic Memory Model: An Evolution of MINERVA2", Proc. Interspeech, 2007.

[8] Hawkins, S., and Smith, R., "Polysp: a polysystemic, phonetically-rich approach to speech understanding". Italian Journal of Linguistics - Rivista di Linguistica, 2001.

[9] Bybee, J., "Phonology and Language Use", Cambridge University Press, 2001.

[10] Tulving, E., "Episodic Memory: from Mind to Brain", Annu. Rev. Psychol. 53, 1-25, 2002.

[11] Luce, P. A. and Lyons, E. A., "Specificity of Memory Representations for Spoken Words", Memory and Cognition, 26(4): 708-715, 1998.

[12] Greenberg, G.Z. and Larkin, W.D. "Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe signal method". Journal of the Acoustical Society of America 44 1513-1523, 1968.

[13] Scharf, B., Quigley, S., Aoki, C., Peachey, N., and Reeves, A. "Focused auditory attention and frequency selectivity". Perception and Psychophysics , 42, 215-223, 1987.

[14] Dai H., Scharf B., and Buus S., 1991. "Effective attenuation of signals in noise under focussed attention.", J. Acoust. Soc. Am. 89 , 1991.

[15] Scharf, B. "Auditory attention: the psychoacoustical approach". In H. Pashler (Ed.), Attention (pp.75-113). Hove: Psychology Press. 1998.

[16] Findlay K, Simpson W, Manahilov V., "Categorical perception requires spatially distributed attention" Perception 31 ECVP Abstract Supplement, 2002.

[17] Langley, P., Laird, J. E., & Rogers, S. "Cognitive architectures: Research issues and challenges". Cognitive Systems Research, in press.

[18] Kolodner, Janet. "Case-Based Reasoning". Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1993.

[19] Aamodt A. and Plaza E, 1994. Case-based reasoning: foundational issues. AI Communications, Vol. 7:1, March 1994.

[20] Anderson, J. R., "The architecture of cognition". Harvard University Press, Cambridge, 1983.

[21] Rouse, W.B and Hurt, R.M. "Human problem solving in fault diagnosis tasks", Georgia Institute of Technology, Center for Man-Machine Systems Research, Research Report no 82-3, 1982.

[22] Kuhl, P. K., "A new view of language acquisition". Proc Natl Acad Sci, 97(22):11850-7, 2000.

[23] Iverson, Paul et al, "A perceptual interference account of acquisition difficulties for non-native phonemes", Cognition, 87(1): B47-B57, 2003