Biologically Inspired Cognitive Architectures 18 (2016) 23–32

Contents lists available at ScienceDirect

Biologically Inspired Cognitive Architectures

journal homepage: www.elsevier.com/locate/bica

Research article

Continuous phone recognition in the Sigma cognitive architecture

Himanshu Joshi [a,b,*], Paul S. Rosenbloom [a,b], Volkan Ustun [a]

[a] Institute for Creative Technologies, University of Southern California, 12015 Waterfront Dr., Playa Vista, CA 90094, USA
[b] Department of Computer Science, University of Southern California, 941 W. 37th Pl., Los Angeles, CA 90089, USA

ARTICLE INFO

Article history:
Received 27 July 2016
Accepted 20 September 2016

Keywords:
Cognitive architecture
Graphical models
Sigma
Speech recognition
Factor graphs
Dynamic Bayesian networks
HMM

ABSTRACT

Spoken language processing is an important capability of human intelligence that has hitherto been unexplored by cognitive architectures. This reflects on both the symbolic and sub-symbolic nature of the speech problem, and the capabilities provided by cognitive architectures to model the latter and its rich interplay with the former. Sigma has been designed to leverage the state-of-the-art hybrid (discrete + continuous) mixed (symbolic + probabilistic) capability of graphical models to provide in a uniform non-modular fashion effective forms of, and integration across, both cognitive and sub-cognitive behavior. In this article, previous work on speaker dependent isolated word recognition has been extended to demonstrate Sigma's feasibility to process a stream of fluent audio and recognize phones, in an online and incremental manner with speaker independence. Phone recognition is an important step in integrating spoken language processing into Sigma. This work also extends the acoustic front-end used in the previous work in service of speaker independence. All of the knowledge used in phone recognition was added supraarchitecturally – i.e. on top of the architecture – without requiring the addition of new mechanisms to the architecture.

Introduction

Cognitive architectures model the fixed structures underlying human intelligence. They are an important approach towards realizing the original goal of AI, a working implementation of a complete cognitive system in aid of creating synthetic agents with human capabilities (Langley, Laird, & Rogers, 2009). The Sigma cognitive architecture is under development at the University of Southern California's Institute for Creative Technologies (ICT) to support the real time needs of intelligent agents – Virtual Humans – that are integrated simulations of human bodies, minds and behaviors in virtual environments, focusing on human-like interactions with real humans. Such virtual humans require a real-time combination of a broad set of human-level capabilities, from central cognitive (thought) processes to peripheral sub-cognitive (perceptual and motor) processes (Campbell et al., 2011). Sigma predicates its ability to support Virtual Humans on the fact that its development is guided by a quartet of desiderata (Rosenbloom, Demski, & Ustun, 2016). Two of the four desiderata derive directly from this requirement of real-time combination of broad capabilities: (1) grand unification, combining both traditional cognitive capabilities and sub-cognitive capabilities such as motor control, vision, and speech; and (2) sufficient efficiency, executing quickly enough for anticipated applications. The third desideratum – (3) functional elegance, generating the broad (cognitive and sub-cognitive) functionality necessary for human-level intelligence from a simple and theoretically elegant base – brings a principle of parsimony into the design of Sigma that encourages uniformity in how disparate capabilities are implemented and thus also potentially enables tight coupling among them. Finally, the remaining desideratum – (4) generic cognition, aims to integrate ideas from both cognitive science and artificial intelligence. This work provides an important step towards grand unification, functional elegance and generic cognition. It does however at the same time raise serious issues concerning sufficient efficiency that will be discussed at the end.

This article builds on previous work on speaker dependent isolated word recognition in Sigma (Joshi, Rosenbloom, & Ustun, 2014). While the focus there was demonstrating a simple form of speech processing, and understanding the extent to which the architecture can contribute towards achieving it, the focus here is on taking the first step towards integrating a speech capability that can process a more natural form of continuous, speaker independent speech signal by segmenting it into a stream of recognized phones. The manner in which this is achieved is both constrained

* Corresponding author at: Institute for Creative Technologies, University of Southern California, 12015 Waterfront Dr., Playa Vista, CA 90094, USA.
E-mail addresses: himanshu@ict.usc.edu (H. Joshi), rosenbloom@ict.usc.edu (P.S. Rosenbloom), ustun@ict.usc.edu (V. Ustun).

http://dx.doi.org/10.1016/j.bica.2016.09.001
2212-683X/

by key aspects of what is known about human speech processing – in particular, the online, incremental nature of it – and uniform with how cognitive processes are implemented within the architecture.

Speech processing involves a combination of challenges – the high dimensional nature of sub-symbolic input, a rich high-bandwidth interplay of sub-symbolic and symbolic aspects of incremental language processing, the task and the goal oriented nature of human conversations, etc. – that have made it a difficult capability to integrate into cognitive architectures. Traditional symbolic cognitive architectures, such as Soar (Laird, 2012), interface to sub-cognitive modules outside of the core architecture for perceptual processing (Laird, 2012; Laird, Kinkade, Mohan & Xu, 2012). In contrast, Sigma's mixed (symbolic + probabilistic) and hybrid (discrete + continuous) core – based on graphical models (Koller & Friedman, 2009) – has already been shown capable of a uniform non-modular integration of symbolic decision making with probabilistic perception in the form of conditional random fields (CRFs) and simultaneous localization and mapping (SLAM) (Chen et al., 2011). More broadly, a variety of forms of memory, learning and decision-making have been demonstrated in Sigma, along with forms of perception and mental imagery (Rosenbloom et al., 2016). The hypothesis underlying this work – and more broadly, the speech processing work in Sigma – is that Sigma is sufficiently capable of implementing speech and providing a tight coupling of it and language processing with cognition in a non-modular fashion on top of the architecture without adding any language or speech specific capabilities to the architecture. As a step towards realizing this vision, this article presents results from a continuous phone recognition task that demonstrates Sigma's ability to segment a stream of continuous audio into constituent phones, and to recognize those phones, in a non-modular supraarchitectural fashion.

Sigma achieves phone segmentation and recognition by leveraging the graphical architecture hypothesis – that the key to making progress on Sigma's desiderata entails a blending of lessons learnt from the history of work on both cognitive architectures and graphical models (Rosenbloom et al., 2016). Although not directly biologically inspired, graphical models share many of the attributes of neural networks and a number of neural network algorithms map directly onto them (Jordan & Sejnowski, 2001). Graphical models have the added advantage of providing a simple path for implementing symbolic processing and integrating it with sub-symbolic processing (Rosenbloom, 2012b; Rosenbloom et al., 2016). Sigma uses factor graphs as its graphical model of choice, with a variant of the summary product algorithm as its solution method (Kschischang, Frey, & Loeliger, 2001). Factor graphs are bipartite, undirected, potentially cyclic, graphical models that are more expressive than Bayesian and Markov networks. They enable grand unification via a mixed hybrid cognitive language, functional elegance through a small set of general mechanisms with broad applicability, and sufficient efficiency and generic cognition through the range of state-of-the-art and cognitively-inspired algorithms that map onto them. The summary product algorithm can be used to compute marginals over all of the variables in the graph, when summarization occurs via sum/integral, or the maximum a posteriori (MAP) distribution, when summarization occurs via max.

Sigma's basis in graphical models enables the core of the phone recognition task – as modeled by hidden Markov models (HMMs) that process low-level spectral speech features derived from an acoustic front-end – to be implemented via the same kinds of knowledge structures it uses for cognitive processing. Furthermore, a portion of the structure of these HMMs can itself be generated automatically by diachronic processing mechanisms previously introduced into Sigma in support of reinforcement learning (RL) (Rosenbloom, 2012a). These results provide an important step towards a uniform, non-modular, speech capability in Sigma that can be coupled tightly with both language and cognition. In the process, these results also provide significant contributions towards both grand unification and functional elegance by yielding these subsymbolic capabilities without additional architectural modules or extensions.

The next section covers background material on the phone recognition task and its role in speech processing within a cognitive architecture. Sections 'Sigma cognitive system and architecture' and 'Continuous phone recognition in Sigma' introduce the most relevant aspects of the Sigma architecture along with how the task at hand maps onto Sigma, together with a discussion of how Sigma automatically generates a portion of the graph used in this task. Section 'Continuous phone recognition in Sigma' then discusses the dataset that was used, the experiments conducted, and the accompanying results. Finally, the article is wrapped up in Section 'Experimental results' with a summary and possible future steps.

Continuous phone recognition and speech processing

Speech is a non-stationary signal produced by the human articulatory apparatus that is assumed to consist of a short-term and a long-term component (Rabiner, 1989). The short-term component is of the order of a few tens of milliseconds (ms) and characterizes the development of individual sounds. The long-term component is of the order of a few hundred milliseconds and characterizes the development of sequences of sounds (Rabiner, 1989). Following the noisy channel model, the audible speech signal is assumed to stem from speech codes that cannot be directly observed. These speech codes can be individual sounds – i.e. *phones*[1] – or contextualized sounds, or even words. Small-vocabulary systems typically work at the level of individual words, whereas medium- or large-vocabulary systems model phones or contextualized phones. Phones model individual sounds and their study – phonology – is based on the assumption that spoken words are composed of smaller units of speech (Jurafsky & Martin, 2008). Choosing to model phones as the acoustic unit of choice allows speech recognizers to use available data effectively and recognize words not present in the training data. From a cognitive linguistics perspective, the parallel architecture hypothesis (Jackendoff, 2007) considers phonology to be an independent generative component of language, alongside structure (syntax) and meaning (semantics). This work focuses on recognizing a stream of spoken audio as a sequence of phones, using a generative model of phones. It serves as a first step towards integrating a full-fledged generative phonology component.

The dominant paradigm in speech recognition involves the use of the generative framework of hidden Markov models (HMMs), with one HMM per speech code, a phone in this case. An HMM consists of a Markov chain that models the evolution of the speech code over time while producing acoustic observations. Each phone is considered to be a finite state automaton (FSA) (Fig. 1), as modeled by the states of the HMM. As each phone automaton progresses, it passes through the beginning, middle and finally the end of the phone while producing the acoustic observations. These observations are what are perceived – i.e. heard – by the human ear, whereas the part of the phone (beginning, middle or end) producing these observations is not observed directly but can be inferred from the acoustic observations. The acoustic observations are obtained via digital signal processing (DSP) that is designed to

[1] Phonology distinguishes between a phoneme – an idealized sound – and a phone, the physical realization of a sound (Jurafsky & Martin, 2008). In this work, phones are of most direct relevance.