Testing a Critique of Computational Cognitive Modelling

Arnaud Henry

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2012

Abstract

This doctoral thesis will present the results of my work into the reanimation of lifeless human tissues.

Acknowledgements

I would like to thanks my awesome flatmates, including the dog, the ghosts and the mice.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Arnaud Henry*)

Table of Contents

Chapter 1

Introduction

Speech recognition is the act of correctly assigning lexical items to an acoustic speech signal. Doing so in an automated way has proven to be a considerable challenge. Although the task seems somehow trivial as we communicate effortlessly with others around us, there are many difficulties arising when one attempts to build a computational model as efficient as we are.

First, each spoken word is composed of speech sounds that are influenced by one another. For example

Secondly, spoken words in a sentence are not usually separated by silence. This means that word boundaries have to be guessed by the model - a problem a written text recogniser does not have to deal with, as written words are separated by blanks. Finally, we often communicate in noisy environments, and each speaker has its own accent and way of pronouncing, not to mention that the intonation of the same utterance by a single speaker changes depending on his mood.

All these facts are a reason as to why no computational model has managed to match our ability to recognise speech.

An approach to solving the problem of automatic speech recognition is the development of computational cognitive models, in which much attention is draw to the cognitive processes involved in speech recognition. These models aim at reproducing empirical data collected through diverse experiments made on human subjects.

1.1 Overview

Chapter 2

Background

2.1 The TRACE model

James McClelland and Jeffery Elman first developed a model of phoneme identification from real speech called TRACE I (Elman and McClelland, 1986). The model had adjustable feature to phoneme connection strengths in order to deal with variability due to local phonetic context. It could extract features from real speech and correctly identify 90% of stop consonants from monosyllable inputs. Because dealing with real speech involved getting concerned with exact duration of different features and phonemes in different contexts, it was harder to focus on the psychological aspect of the model and thus was not extendible in that form. This is why the authors developed another model, TRACE II, in which input to the model was mock speech, but where a lexical layer of processing could be added, enabling to account for a large amount of available psychological data. This is the model we are interested in and which is simply referred to TRACE in the literature.

2.1.1 Motivations

The interactive activation architecture of the TRACE model, which we will describe in detail in the next section, is motivated by several aspects of speech that have been discussed by Klatt (Klatt, 1979) and which McClelland and Elman refer to in their seminal paper:

- Speech is spread across time. This fact implies that a model of speech recognition has to keep a record of previous utterances in order to process incoming input. Not only previous context, but it has been shown that also subsequent

context has an influence on perception at a specific time (Salasoo and Pisoni, 1985) (Thompson, 1984). This means that a successful model cannot have static phoneme or word recognition, in the sense that it must be possible to alter the perception of previously recognised entities.

- Speech units overlap. Whether at the word or phoneme level, there are no reliable cues as to where units are separated. Indeed, words in fast speech - that is, conversational speech - tend to run into one other, sometimes sharing a phonetic sound in order to make the pronunciation easier for the speaker. In a similar way, phonemes overlap in such a way that it is impossible to tell at which point in time each phoneme starts and ends. In contrast to these two observations, written text has an advantage as letters are clearly distinguishable and words are separated by white spaces. The fact that speech units overlap means that phoneme and word recognition doesn't rely on precise segmentation of the input.

- Speech units are context-sensitive. The way phonemes are pronounced depends greatly on surrounding phonemes. For example,

  This effect known as coarticulation has caused many problems to cognitive modellers. McClelland and Elman take it into account by allowing connections to alter depending on the context.

- Speech is perceived in noise. We often converse rather efficiently in noisy environments, and this fact alone implies that a model of speech recognition must not rely on the accurate recognition of any part of the speech.

Most of these facts are peculiar to speech and make the general task of speech recognition more difficult than written text recognition.

2.1.2 Architecture

The TRACE model is composed of three layers of processing: features, phonemes and words, in which every unit is replicated across time.

At the feature level, a set of seven dimensions is used to represent mock speech: Consonantal, Vocalic, Diffuseness, Acuteness, Voicing, Power and Burst. The first five were taken from the famous work of (**?**), Power was included to differentiate further vowels from consonants, and Burst was introduced to help discriminate among the different stop consonants. Each of the seven dimensions is divided into nine value

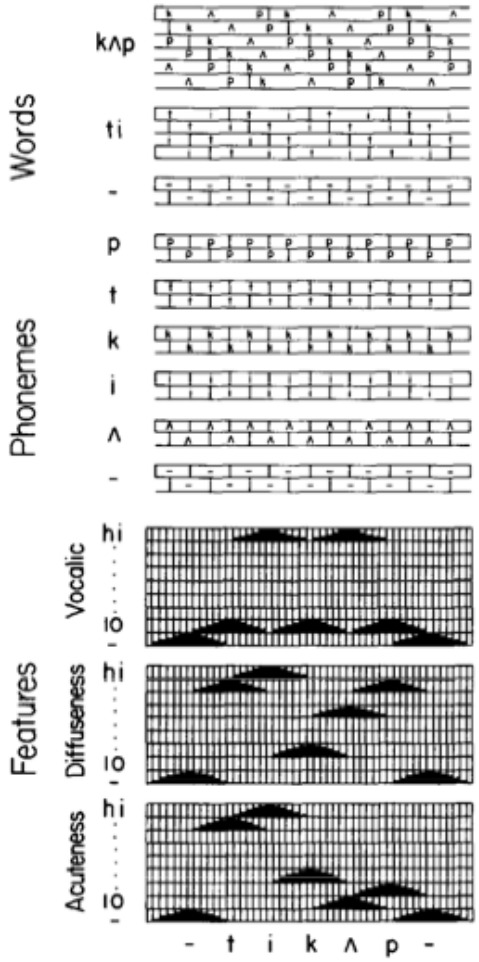

Figure 2.1: A subset of the units in TRACE. Each rectangle represents a different unit at each level. The input feature specifications for the phrase tea cup are indicated for the three following dimensions: Acuteness, Diffuseness and Vocalic. (McClelland and Elman, 1986)

ranges, eight of which correspond to a grade from low to high. The ninth value is reserved to represent silence, and thus is used only by the silence phoneme ”-”. The 63 (7*9) units are replicated for each of the 25ms time slices of the model, and are filled in progressively by values corresponding to the input of the model. Figure 2.1 shows an example of a complete filling for three of the seven features for the input *tea cup*.

At the phoneme level, 15 phonemes are represented by units spanning 6 time slices and replicated every 3 time slices, as can be seen in figure 2.1 for a few phonemes. Note that this means adjacent phoneme units overlap, just as it is observed in real speech. Each phoneme unit receives activation from the feature units that make up that phoneme. McClelland and Elman came up with feature specifications for the following

14 phonemes: /b/, /p/, /d/, /t/, /g/, /k/ (stop consonants), /s/, /S/ (fricatives), /l/, /r/ (liquids), /a/, /i/, /u/ and /Î/ (vowels), and also added a silence phoneme /-/. There are inhibitory connections between units representing different phonemes which overlap in time, such that if a certain phoneme at a certain time slice receives activation from the feature level, it will inhibit other phonemes units at that time slices and three time slices apart (since a phoneme unit spans 6 time slices).

At the word level, each word is represented by units replicated every 3 time slices (as for the phonemes). The time span of these units depend on the length of the word, 6 time slices being allocated to each phoneme which compose that word (referring to the example in figure 2.1, /kp̂/ spans 18 time slices and /ti/ spans 12 time slices). Word units receive activation from the phoneme units that match and are aligned with the particular phonemes that make up the word. For example, the phoneme /t/ at time slice 6 will activate all word units starting at that time with /t/ being their first phoneme, and also all words units starting at time slice 0 which have /t/ in second position. As for the phoneme level, word units inhibit one another if they represent a different word and are overlapping in time. This is the basis of the interactive activation mechanism present in TRACE.

In addition to the excitatory bottom-up connections between layers and intra-layer inhibitory connections, TRACE provides top-down excitatory connections from words to phonemes. Each phoneme unit can therefore receive activation from words which contain that phoneme at that particular time slice.

2.1.3 Advantages

2.1.4 Limitations

2.2 Evolution of cognitive modelling

(Chawla and Shillcock, 2011)

Chapter 3

Methodology

3.1 Philosophical critique

3.2

Check out figure 3.1

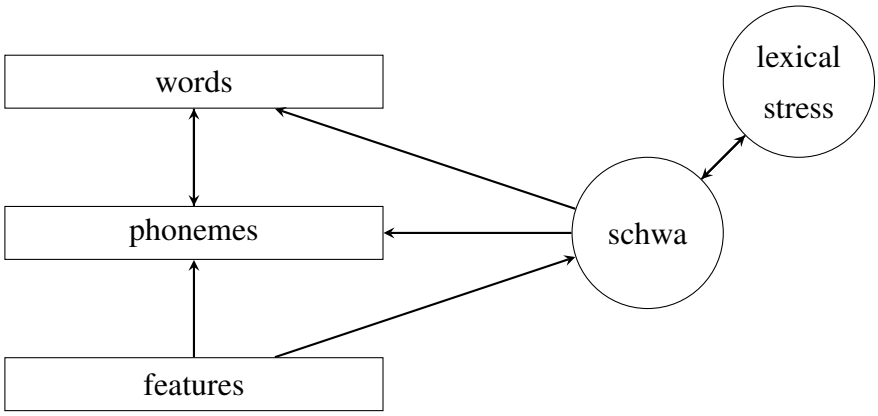

Figure 3.1: Architecture of the modified model

Chapter 4

Evaluation

4.1 Hypotheses

4.2 Dataset

4.3 Experiments

Chapter 5

Conclusion

5.1 Future Work

Bibliography

Chawla, M. and Shillcock, R. (2011). The development of a successful cognitive model; a case-history of the trace model of speech perception. *in Press*.

Elman, J. and McClelland, J. (1986). Exploiting lawful variability in the speech wave. *Invariance and variability in speech processes*, pages 360–385.

Klatt, D. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, 7(312):1–26.

McClelland, J. and Elman, J. (1986). The trace model of speech perception. *Cognitive psychology*, 18(1):1–86.

Salasoo, A. and Pisoni, D. (1985). Interaction of knowledge sources in spoken word identification. *Journal of Memory and Language*, 24(2):210–231.

Thompson, H. (1984). Word recognition: A paradigm case in computational (psycho-)linguistics. In *Proceedings of the Sixth Annual Meeting of the Cognitive Science Society, Boulder, CO*.