A phonologically motivated input representation for the modelling of auditory word perception in continuous speech

Richard Shillcock
Centre for Cognitive Science
University of Edinburgh
2 Buccleuch Place
Edinburgh
rcs@cogsci.ed.ac.uk

Geoff Lindsey
Dept. of Linguistics
University of Edinburgh
AFB, George Square
Edinburgh
geoff@ed.ling.ac.uk

Joe Levy
HCRC
Univ. of Edinburgh
2 Buccleuch Place
Edinburgh
joe@cogsci.ed.ac.uk

Nick Chater
Dept of Psychology
Univ. of Edinburgh
7 George Square
Edinburgh
nicholas@cogsci.ed.ac.uk

rcs@cogsci.ed.ac.uk@nsfnet-relay.ac.uk

Abstract

Representational choices are crucial to the success of connectionist modelling. Most previous models of auditory word perception in continuous speech have relied upon a traditional Chomsky-Halle style inventory of features; many have also postulated a localist phonemic level of representation mediating a featural and a lexical level. A different immediate representation of the speech input is proposed, motivated by current developments in phonological theory, namely Government Phonology. The proposed input representation consists of nine elements with physical correlates. A model of speech perception employing this input representation is described. Successive bundles of elements arrive across time at the input. Each is mapped, by means of recurrent connections, onto a window representing the current bundle and a context consisting of three such bundles either side of the current bundle. Simulations demonstrate the viability of the proposed input representation. A simulation of the compensation for coarticulation effect (Elman and McClelland, 1989) demonstrates an interpretation which does not involve top-down interaction between lexical and lower levels. The model described is envisaged as part of a wider model of language processing incorporating semantic and orthographic levels of representation, with no local lexical entries. [1]

Introduction

Psychologists wishing to model spoken language perception have typically assumed that the physical speech signal may be translated into a featural level of representation, which then maps onto a phonemic level, which, in turn, supplies activation to a lexical level. In the TRACE model (McClelland & Elman 1986) these three levels of representation are instantiated in an interactive-activation architecture, in which patterns of activation percolate up and down between adjacent levels. In models of speech production, similar assumptions are made. In the Plaut and Shallice (*in press*) model of pronunciation in deep dyslexics, patterns of activity at the semantic level are mapped onto position-specific phonemic representations. There have been some departures, however, from explicit, local phonemic representations. In the Seidenberg and McClelland (1989) model of word naming, orthographic patterns are mapped onto Wickelphones consisting of triples of consecutive features. In the model of word recognition proposed by Norris (1990), bundles of distinctive features arrive across time at the 11 input nodes and activation is mapped onto an output layer consisting of local representations of words.

In some of these models, the phonological representations employed, although internally consistent, are simplifications of what might be thought adequate by phonologists. In others, while the input representations are relatively sophisticated, they do not reflect phonologists' more recent disenchantment with the phoneme and with bundles of features organized strictly linearly. From a different perspective, many researchers in automatic speech recognition (ASR) have become disillusioned with the notion that ASR is best attempted by recognizing SPE-style features (Chomsky & Halle, 1968) in the physical signal and subsequently parsing them into phonemes and words.

Although the models cited above have achieved considerable success in capturing many of the qualitative aspects of language processing, the enterprise should be substantially improved by the use of a speech input representation which, first, is consonant with current phonological theory and, second, has a consistent relationship with the physical speech signal. Such an input representation will more

[1] This research was carried out under E.S.R.C. grant number R000 23 3649.

adequately reflect the structure of the real-world problem of speech recognition. Below we present such an input representation – an alternative to the orthodox SPE-style framework – and describe its instantiation in a recurrent network. We then review its success as a psychological model.

Phonological motivation

The input representation described in Table 1 is based on recent work in Government Phonology (Kaye, Lowenstamm & Vergnaud, 1985, 1990). The speech signal is decomposed into nine elements, defined briefly as follows.

A: oral cavity openness; alone, the vowel quality of *palm*.
I: palatality; alone, the vowel quality of *see*.
U: labiality; alone, the vowel quality of *boot*.
?: occlusion; abruptness; alone, glottal stop.
h: aperiodic energy; alone, [h].
N: nasality.
R: apicality/coronality/coronal formant locus.
@: velarity/centrality.
H: voicelessness.

The elements are represented principally in a binary way. In four cases a value of 0.5 is used; this is a representational compromise which reflects the notion of "government" within the phonological theory. The lefthand column gives Machine-Readable Phonetic Alphabet (MRPA) equivalents for short vowels and syllable-initial consonants; elements may be subtracted from the definitions in Table 1 to represent segments in other environments (*e.g.* /k/ would lack the element **h** when unreleased as in *act*). In Table 1, the initial glides in *yet* and *wet* have the same element representations as

The input representation

	elements								
segment	?	h	U	N	R	@	H	I	A
p (pat)	1	1	1	0	0	0	1	0	0
t (tap)	1	1	0	0	1	0	1	0	0
k (cat)	1	1	0	0	0	1	1	0	0
b (bat)	1	1	1	0	0	0	0	0	0
d (dot)	1	1	0	0	1	0	0	0	0
g (got)	1	1	0	0	0	1	0	0	0
m (mill)	1	0	1	1	0	0	0	0	0
n (nil)	1	0	0	1	1	0	0	0	0
ng (sing)	1	0	0	1	0	1	0	0	0
f (fit)	0	1	1	0	0	0	1	0	0
th (thin)	0	.5	0	0	1	0	1	0	0
s (sin)	0	1	0	0	1	0	1	0	0
sh (shin)	0	1	0	0	1	0	1	1	0
zh (measure)	0	1	0	0	1	0	0	1	0
h (hat)	0	1	0	0	0	0	0	0	0
v (vat)	0	1	1	0	0	0	0	0	0
dh (that)	0	.5	0	0	1	0	0	0	0
z (zen)	0	1	0	0	1	0	0	0	0
l (lip)	1	0	0	0	1	0	0	0	0
r (rip)	0	0	0	0	1	0	0	0	0
y (yell)	0	0	0	0	0	0	0	1	0
w (well)	0	0	1	0	0	0	0	0	0
i (bin)	0	0	0	0	0	0	0	1	0
e (den)	0	0	0	0	0	0	0	1	.5
a (ban)	0	0	0	0	0	0	0	.5	1
o (don)	0	0	1	0	0	0	0	0	1
uh (bud)	0	0	0	0	0	1	0	0	1
u (wood)	0	0	1	0	0	0	0	0	0
@ (about)	0	0	0	0	0	1	0	0	0
A (see text)	0	0	0	0	0	0	0	0	1

Table 1. Definition of short vowels and syllable-initial consonants of standard Southern British English in terms of elements. The special element A does not appear singly in isolation (see Table 2, overpage).

segment	expansion
ch (chair)	t sh
jh (journey)	d zh
@@ (bird)	@ @
ou (bode)	@ u
oi (boy)	o i
oo (bored)	o o
ii (bee)	i i
uu (boot)	u u
ei (bade)	e i
ai (bide)	A i
au (loud)	A u
u@ (poor)	u @
i@ (beer)	i @
e@ (pair)	e @
aa (bard)	A A

Table 2. Expansions of the 2 affricates and the 13 diphthongs and long monopthongs of standard Southern British English.

the vowels of *pit* and *put*, respectively; the differences are attributed to location in syllable structure and are not explicitly encoded in the model described below. Affricates, diphthongs and long monopthongs are decomposed into two consecutive segments, as shown in Table 2.

The model

The goal is a comprehensive model of speech processing which will allow the detailed modelling of psycholinguistic data, going beyond the essentially qualitative results achievable with models such as TRACE. The earliest point at which to begin the psychological modelling of speech processing is with the transduction of the physical signal into some common currency of activation. A noisy acoustic signal is converted into a representation which has psychological significance. Sensitivity to a window of context is essential if this conversion is to be reliable. Some parts of the signal will be captured more securely than others, either because they are inherently more distinctive or because of context. The input representation detailed above is seen as the most appropriate description of the input at this point.

The approach taken below is defined, first, by the fact that, although the elements have a relationship with the physical speech signal, there is no discrete set of acoustic entities which might be employed in a mapping from an element level of representation. Still less is there available a corpus of natural speech transcribed in acoustic terms. The second constraint on the modelling is the principle of eschewing intermediate levels of linguistic representation, phonemes in particular. In this approach, the aim is to develop a single, psychologically realistic level of representation of the speech stream and to map this directly to a level at which the semantics of individual words is expressed. (This approach postpones segmentation and binding issues.)

In accordance with these constraints, the model described below captures the context-dependence of the initial transduction of the speech signal by auto-associating the patterns of elements as they arrive at successive time-slices. We envisage a subsequent mapping onto semantic representations; only the initial auto-association is described here, however. It is predicted that certain processing, typically ascribed to higher (lexical, morphological) levels of description, will be more or less weakly foreshadowed at this low level. (Norris (*in press*) demonstrates that, in principle, connectionist models which learn are likely to encode aspects of higher-level generalizations at lower levels of representation, giving rise to behaviour which looks like traditional "top-down" interaction.) The model we describe, once trained, accepts transcribed stimulus materials from experiments on spoken word recognition and outputs quantitative data on the facility with which each segment may be identified and represented.

Previously some of the present authors have suggested that some aspects of auditory word perception can be captured in a mapping between a featural level of representation and a phonemic level, in a recurrent network (Shillcock, Levy and Chater, 1991; Levy, Shillcock and Chater, 1991). This was an attempt to model speech processing in a comprehensive, full-scale way which postponed incorporating any local representations of lexical entries. The model was motivated by the Seidenberg and McClelland (1989) model of word naming, in which psychologically interesting behaviour falls out of a simple mapping between orthography and phonology, and in which a "lexical entry" is a distributed entity. Our goal was to remain within the auditory modality and to assess the extent to which behaviour previously attributed to higher-level (morphological and lexical) representations could be captured at the lower, featural and phonemic levels. This model illustrated Norris's observation about lower levels of representation embodying aspects of higher-level generalizations. Thus, for instance, the model correctly predicted five out of the six phoneme restorations reported by Elman and McClelland (1989) in their demonstration of compensation for coarticulation being triggered by a restored phoneme. The model was able to achieve the same sort of restoration on the basis of a learned features-to-phonemes mapping, with no explicit representations of words.

The new input representation described above allows us to continue with the investigation of a mapping between observable and necessary levels of representation (orthography, phonology, semantics) while avoiding local representations, at intermediate levels, of less defensible categories (phonemes, local lexical entries). Accordingly, the earlier model of

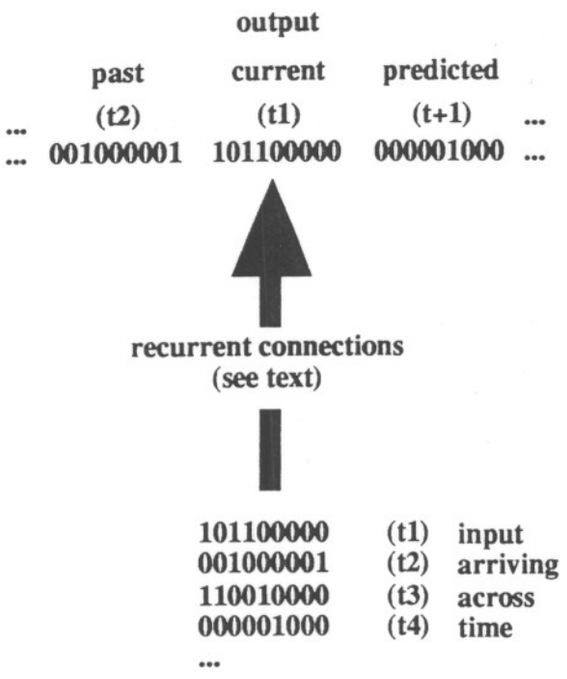

Figure 1. Input across time is mapped onto a more stable representation.

speech processing is superseded by that shown in Fig. 1.

The mapping currently implemented between the two representations is a "cut-down" version of "back-propagation through time", unfolding the network once rather than many times (Servans-Schreiber, Cleeremans & McClelland 1989; Chater 1989; Chater & Conkey, *submitted*) and thus sacrificing some of the ability reliably to pick up long distance dependencies, in exchange for speed of training. This "copyback" structure was introduced by Elman (1988, 1990) and Norris (1988). It will be superseded, in the present model, by a full "back-propagation through time" algorithm (Rumelhart, Hinton & Williams, 1986), in which a recurrent network is unfolded into many copies arranged in a feed-forward architecture, with standard back-propagation being applied to the result.

There are nine input nodes corresponding to the nine elements described above. There are thirty hidden units. The output nodes are grouped into seven sets of nine representing the bundles of elements at particular time-steps. The exact choice of the number of time-steps represented in the output is arbitrary, to an extent, and at this stage is motivated by the desire to study the effects of phonotactic constraints. (In practice, most of the observation of the behaviour of the model has involved recording the activation pattern for the current time-step.) A window which stretches over several segments allows right-context effects to be studied. In reality, the effects due to coarticulation with the immediately adjacent time-step are the most important, although vowel-to-vowel effects require a window of several time-steps. Note that including the surrounding feature-bundles in the output layer forces the network to learn the context in which any current feature-bundle occurs.

Although current simulations have used the idealization of one segment, or bundle of elements, to each time-step, it is possible to relax this aspect of the model by, for instance, centring each bundle of elements around three consecutive time-steps. Elements may then be allowed to spread into adjacent time-slices occupied predominantly by the adjacent bundle of elements. Thus, the nasalization of the vowel in *don* may be represented by the element N being present in one or all of the time-slices occupied by the elements corresponding to the vowel. This move would also facilitate the vexed issue of the representation of diphthongs and long monopthongs.

Training the network

The current version of the model employs the "copyback" structure described above. In the simulations reported below, a learning rate of 0.1 was employed; momentum was not used. The network was trained until it began to show signs of overfitting – training that resulted in a decrease of error for the training set but led to increasing error for a separate test set was disregarded. This required between 500 and 600 epochs. To encourage the network to employ context, noise was added to the input; there was an 11% probability that any element in the input would have its value changed to or from 0. The noise was generated on-line and was different for every epoch of training. The learning phase of the simulations was quite computationally intensive, using 30-40 CPU hours on a variety of SUN SPARC-based machines, using a customized version of the Rumelhart and McClelland (1988) simulation package.

The initial, limited training data was derived from some 3490 words worth of spoken discourse, taken largely from the LUND Corpus (Svartik & Quirk 1980). This corpus consists of a word level transcription and includes filled pauses, false starts and corrections. This was converted automatically to an idealized segment-level transcription which was, in turn, converted to the nine elements described above. The training set was made up of 9097 segments and a test set of 3285 segments was used to test for overfitting. No attempt was made to impose phonological reduction or coarticulation.

Modelling psycholinguistic data

The success of the model in representing a particular bundle of elements in the output level was measured by calculating the sum of squares of the error associated with that bundle when compared with each

of the possible input patterns. This allowed us to determine, for instance, whether a particular output resembled the expected pattern for /s/ or /sh/. Initial simulations illustrate the potential of the model and of the wider approach.

Sensitivity to context

The model is sensitive to segmental context. Auto-associated patterns of elements in "current" position are more accurate, in terms of sums of squares of the error, when the model is given input from transcribed normal discourse than when the same bundles of elements are presented in random order. The model relies on previous context to identify the current bundle of elements; when this context is aberrant, it hinders correct recognition.

Human listeners employ context both before and after the segment in question. Training with noise forced the network to rely on both "left" and "right" context. The scores for the bundles of elements in "past" positions for normal and abnormal discourse indicated that the model is sensitive to right context in recognizing bundles of elements, and was misled by abnormal right context.

Phoneme restoration and compensation for coarticulation

Listeners' perception of degraded individual speech sounds in words is often restored (Warren 1970), particularly when the intended phoneme and the replacing sound (*e.g.* white noise, a click, silence) are similar, and when replacement occurs after the uniqueness point of the word.

This effect is often not compelling, however, and there are inherent difficulties in interpreting the effect. An important exception to this latter problem is the demonstration of compensation for coarticulation reported by Elman and McClelland (1989). This particular experiment is of crucial theoretical interest because of the claim that it demonstrates top-down lexical influences on lower-level phonemic representations. It is the strongest experimental evidence for phoneme restoration, and for top-down influences on perception in general. Most of the six words employed in that study (*Christmas, copious, ridiculous, foolish, English, Spanish*) end with suffixes, suggesting that the phoneme restoration reported for the final segment might emerge from a model which only encoded low-level statistical generalizations about spoken English. Suffixes are frequent sequences of segments and such a model might simply encode the knowledge that the sequences corresponding to *-ish* and *-ous* are more likely in some contexts than in others, without having anything like an adequate representation of morphological categories.

The model was given transcriptions of the words listed above, in neutral left-contexts, with the final segment replaced in each case by an identical hybrid segment intermediate between /s/ and /sh/. (/s/ and /sh/ differ only on the palatality, I, element; this was replaced by 0.5 to create the intermediate segment.) The scores assigned to the critical segment in "current" position were recorded and the sum of squares of the error calculated for the bundles of elements corresponding to /s/ and /sh/ respectively. The model has an overall preference for the /s/ interpretation, reflecting the relative preponderance of /s/ over /sh/ in the training corpus. Crucially, however, when the difference between the two sums of squares, for /s/ and /sh/ for each word, is calculated the model exhibits precisely the pattern of restoration found in human subjects. In Fig. 2. "preference for /s/" is the difference between the two sums of squares. It would therefore be possible for a categorical perception criterion to be placed on the /s/–/sh/ continuum so as to ensure that appropriate phoneme restoration occurs for each word.

This simulation suggests that there is no need to posit top-down interaction, as traditionally conceived, to explain Elman and McClelland's demonstration of phoneme restoration.

Conclusions

The proposed input representation, based on the elements of Government Phonology, is a viable alternative to ones comprising SPE-style features, for the connectionist modelling of speech processing. The input representation used gave simulation results closer to the human data than did the previous input based on SPE-style features. Simulation of the compensation for coarticulation effect suggests that this effect, which was previously interpreted as a top-down lexical effect, may be the result of learning simple statistical generalizations within speech input. Apparent "higher-level" generalizations are apparent, to differing degrees, at very low levels. A more accurate view of what is possible at such an early stage limits what it is necessary to explain at putative higher levels.

Increasing the training corpus from the current 3490 word tokens (905 word types) will improve the performance of the model. Introducing into the training corpus plausible levels of coarticulation (by means of element spreading) and phonological reduction (by means of element and segment elimination) will improve the performance by adapting the context of any particular element or bundle of elements in a predictable way.

Many psycholinguistic phenomena which have been taken to involve access to specific representations of spoken words may be explained in terms of the low-level statistical structure of the speech input, as encoded in connectionist models of the process. There

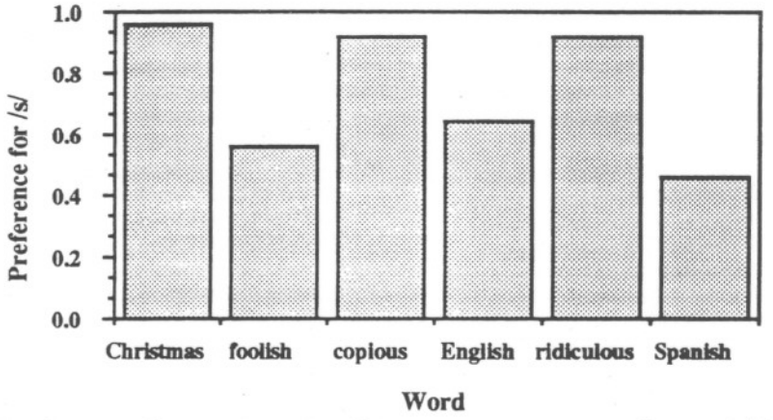

Figure 2. Preference for the bundle of elements corresponding to /s/ is greatest in the appropriate words.

is a methodological imperative within psycholinguistic research to allow "higher level" interpretation of data only when low level explanations can be ruled out.

References

Chater, N. 1989. Learning to respond to structures in time. Technical Report RIPRREP/1000/62/89, RSRE, Malvern, Worcs.

Chater, N, & Conkey, P. (*submitted*). Finding linguistic structure with recurrent networks.

Chomsky, N. & Halle, M. (1968). *The Sound Pattern of English*. Harper & Row, New York.

Elman, J. L. 1988. Finding structure in time. Technical Report, CRL TR 8801, Centre for Research in Language, UCSD.

Elman, J. L. 1990. Finding structure in time. *Cognitive Science*, 14:179-211.

Jakobson, R., G. Fant and M. Halle 1952. *Preliminaries to Speech Analysis*. Technical Report 13, M.I.T. Acoustics Laboratory, Mit Press.

Kaye, J. D., Lowenstamm, J. & Vergnaud, J. -R. (1985) The internal structure of phonological elements: a theory of charm and government. *Phonology Yearbook* 2, 305-328.

Kaye, J.D., Lowenstamm, J. & Vergnaud, J.-R. (1990) Constituent structure and phonological government. *Phonology* 7.

Levy, J., Shillcock, R.C. & Chater, N. (1991). Connectionist modelling of phonotactic constraints in word recognition. Proceedings of the *IJCNN*, Singapore, 1991.

McClelland, J. L. & Elman J. L. 1986. Interactive processes in speech perception: the TRACE model. In D. E. Rumelhart & J. L. McClelland eds. *Parallel Distributed Processing*, Vol. 2., 58-121, Cambridge, Mass: MIT Press.

McClelland, J. L. & Rumelhart, D. E. 1988. *Explorations in Parallel Distributed Processing: Models, Programs and Exercises*. Cambridge, Mass: MIT Press.

Norris, D. G. 1990. A dynamic-net model of human speech recognition. In (G. Altmann, ed.) *Cognitive Models of Speech Processing: Psycholinguistic and cognitive perspectives*, MIT Press.

Norris, D. G. (*in press*). Bottom-up connectionist models of "interaction". To appear in (G. Altmann & R. Shillcock, eds.) *Cognitive Models of Speech Processing: 2nd Sperlonga workshop.*

Plaut, D.C. & Shallice, T. (1991). Deep dyslexia: A case study of connectionist neuropsychology. *Ms.*

Seidenberg, M. S. & McClelland, J. L. 1989. A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523-568.

Servans-Schreiber, D., Cleeremans, A. & McClelland, J. L. 1989. Learning sequential structure in simple recurrent networks in D. Touretsky ed. *Advances in Neural Information Processing Systems*, Vol 1, Morgan Kaufman, Palo Alto, 643-653.

Shillcock, R.C., Levy, J., & Chater, N. (1991). A connectionist model of auditory word recognition in continuous speech. *Proceedings of the Cognitive Science Society Conference*, pp. 340-345, Chicago, 1991.

Svartvik, J., & Quirk, R. 1980. *A Corpus of English Conversation*. Lund: Gleerup.

Warren, R. M. 1970. Perceptual restoration of missing speech sounds. *Science*, 167, 392-393.