

A Connectionist Model for Bootstrap Learning of Syllabic Structure

Jean Vroomen

Department of Psychology, Tilburg University, The Netherlands

Antal van den Bosch

*Department of Computer Science/MATRIKS, Universiteit Maastricht,
The Netherlands*

Beatrice de Gelder

Department of Psychology, Tilburg University, The Netherlands

We report on a series of experiments with simple recurrent networks (SRNs) solving phoneme prediction in continuous phonemic data. The purpose of the experiments is to investigate whether the network output could function as a source for syllable boundary detection. We show that this is possible, using a generalisation of the network resembling the linguistic *sonority principle*. We argue that the primary generalisation of the network, that is, the fact that sonority varies in a hat-shaped way across phonemic strings, ending and starting at syllable boundaries, is an indication that sonority might be a major cue in discovering the essential building bricks of language when confronted with unsegmented running speech. The segment which is most directly related to sonority patterns, the syllable, has received considerable attention in psycholinguistics as being an element of natural language that is easily grasped by language learners. The phoneme prediction network presents a simulation of the necessary *bootstrap* to arrive at the discovery of syllabic segmentation in unsegmented speech, which can be used as a basis for the segmentation of larger structures like words.

Requests for reprints should be addressed to Dr J. Vroomen, Dept. of Psychology, Faculty of Social Sciences, Tilburg University, PO Box 90152, NL-5000 LE Tilburg, The Netherlands.

Thanks are due to Walter Daelemans and Eric Postma for discussions, comments, and support, to the anonymous LCP reviewers for valuable suggestions, and to Geer Hoppenbrouwers and Ton Weijters for transcribing the Dutch data and kindly granting us permission to use the data in our experiments.

Antal van den Bosch is now at the Department of Computational Linguistics, Tilburg University.

INTRODUCTION

This article examines a possible solution to the problem of segmentation of continuous speech. By means of a series of simulations, we investigate whether a connectionist network can provide useful cues for the segmentation of a continuous phonemic speech sequence into meaningful units.

The task of a listener is to map the sound of an utterance to meaning. In many languages, speakers do not provide reliable acoustic cues to the boundaries of words. Nevertheless, listeners have to divide the speech signal into units which can be mapped to sublexical or lexical entries. Several proposals have been formulated to partition and align the speech signal in a way so that only a relatively small set of lexical candidates is entertained. Probably the best known model of spoken word recognition is Marslen-Wilson's (1987) Cohort model. In Cohort, the initial part of the word activates a cohort of lexical hypotheses. Words are dropped out from the cohort until one word becomes unique. At that stage, the word is recognised. This activation/selection process works well with long words presented in isolation, but Cohort does not provide a solution to segmentation of continuous speech. The model presupposes that the boundaries of a word are known, because that is the place where activation starts to build up. However, the model will fail if more than two words have to be recognised of which the first one is not uniquely specified at its offset. This is most likely to occur with short words. In that case, the first word will not be recognised because other candidates are still entertained, and, consequently, the boundary of the second word will be mislocated. For that reason, Cohort would run into serious problems if it were applied to speech recognition of concatenated words.

More attention to the segmentation problem has been devoted in the TRACE model of McClelland and Elman (1986). Segmentation in TRACE is accomplished at the lexical level via the interactive activation and inhibition of competing word candidates. The model is, in general, successful in identifying a word boundary between two concatenated words, but whether it can deal with more than two words is at present unclear (for a detailed discussion, see Frauenfelder & Peeters, 1990).

The metrical segmentation strategy (MSS), as advanced by Cutler and Norris (1988), deals more specifically with continuous speech segmentation. The idea is that listeners exploit the lexical statistics of the language. For English, these statistics are such that content words are likely to begin with strong syllables. According to the MSS, listeners exploit these statistics by segmenting the speech signal whenever a metrical strong syllable is encountered. Generally speaking, the MSS assigns the rhythmic unit of a language a special status as segmentation unit. For English, this unit is

stress-based, but there is also evidence that for a syllable-timed language like French, it is the syllable that acts as a segmentation unit. An empirical demonstration of the latter was found by Mehler, Dommergues, Frauenfelder, and Segui (1981). They observed that French subjects responded faster in a segment-monitoring when the target matched exactly the syllable of a word, rather than when it comprised more or less of the syllable. In a stress-timed language such as Dutch, there is also evidence for a role of the syllable in speech segmentation. For example, Vroomen and de Gelder (1997) observed that words embedded at the end of other words were temporarily activated if their onset matched the beginning of a syllable (e.g. *boos*, meaning “angry”, as embedded in *framboos*, meaning “raspberry”). In contrast, there was no activation of the embedded word if its onset did not match the onset of a syllable, such as, for example, *wijn* (meaning “wine”) as embedded in *zwijn* (meaning “swine”). Embedded words were thus only activated if their onset matched the beginning of a syllable.

From a developmental standpoint, however, all the previously mentioned proposals are associated with serious problems. Segmentation hinges on either lexical knowledge (Cohort and TRACE), or at least on knowledge of the statistical properties of the lexicon (MSS). The newborn’s task, however, is to build a lexicon from scratch, and it thus still needs to acquire such a knowledge base. The models mentioned so far leave therefore unexplained how the child, who does not know the words of the language, segments the speech signal. On what basis does it start? If segmentation indeed hinges on lexical knowledge, then one needs a kind of bootstrap procedure.

One source of information in the speech signal that might serve as a platform for acquiring a lexicon is the fact that there is always syllabic structure in the sound sequences of the language. Words generally start at the onset of a syllable (unless they are resyllabified), and discovering syllabic structure could therefore be of help for discovering word boundaries. The question we address in this article is whether this syllabic structure can indeed be discovered from a raw unsegmented sound sequence.

Before outlining our question, some terminology needs to be defined concerning the notion of *syllable* and *sonority*. For Dutch, Trommelen (1981) proposes a tree-structure representation of the syllable, on the basis of which we present the basic syllable tree representation for Dutch in Fig. 1. This basic structure validates syllable structures such as V as in the first syllable of *adelaar* (“eagle”), /a/; CV as in the monosyllabic word *na* (“after”), /na/; CVC as in *bed* (“bed”), /bet/; and CCV as in the first syllable of *breken* (“to break”), /bre/. Alternatively, this basic structure

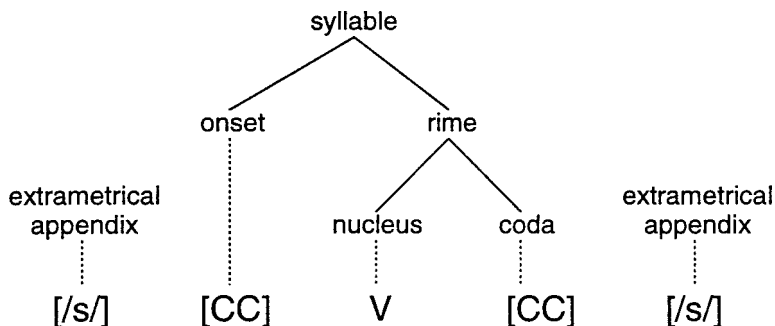


FIG. 1. Tree-structure representation of the Dutch syllable as proposed by Trommelen (1981).

prohibits syllabic structures such as *C¹, *CCC, or *VCCCCC. In Dutch, a syllable consists of (a) an optional *onset* which can contain a sequence of up to two consonants (C); (b) a *nucleus*, the only obligatory element of a Dutch syllable, which contains a vowel (V) or a diphthong; and (c) a *coda* which can contain a sequence of up to two consonants. However, both three-consonant onsets (as in /strat/, *straat*, “street”) and three-consonant codas (as in /bɔrst/, *borst*, “breast”) occur in Dutch. To account for this, metrical phonological theory asserts that the initial /s/ in three-consonant onsets (usually starting with /st/ or /sʃ/) and the final /s/ in three-consonant codas (usually ending with /ts/) are regarded as extrametrical appendices (Halle & Vergnaud, 1980; Trommelen, 1981). The reason for this apparently complex solution to assume extrametrical phonemes rather than just assuming three optional consonants in both the onset and the coda is rooted in the assumption of the *sonority principle*.

We adopt the notion of *sonority* from Selkirk’s (1984) definition of the sonority principle (SP) (but see also Kiparski, 1979, for a similar account). The SP states that within a syllable, sonority starts low at the onset, increases towards a peak value at the nucleus position (hence the name *peak* which is sometimes used as an alternative for *nucleus* in the literature), and gradually decreases along the coda until the end of the syllable. The concept of sonority, though, is not well-defined. In speech-acoustical terminology, sonority is referred to as the ratio of the volume and the effort, but it is also referred to as the degree of openness of the vocal tract (Selkirk, 1984). Measuring this openness during the pronunciation of phonemes by a large number of speakers renders average sonority values per phoneme. When referring to sonority as an phonemic feature, however, it is often taken to be a binary or ternary value (e.g. low, mid, and high sonority). As a working

¹When a presented example is incorrect in the sense that it will not occur in the language (in our case, Dutch), it is preceded by a superscript asterisk (*).

definition, we adopt Selkirk's (1984) proposal. Here, the sonority scale is a compromise of both opinions, by grouping the existing set of phonemes of English into 10 discrete classes. These classes are assigned sonority values between 0.1 (representing voiceless stops) and 1.0 (representing /a/-like vowels), with steps of 0.1. In Table 1 the 10 classes are listed with a short description, the phonemes that occur in our data set and belong to one class, and the sonority value assigned to that class. In grouping the phonemes in our data set into Selkirk's (1984) classification of the sonority of phonemes, we have transferred a classification that was originally illustrated to divide English phonemes to Dutch. None the less, the classes are claimed to be at least fairly universal for all languages. Moreover, Dutch phonology is relatively close to English phonology. We therefore take Table 1 to provide a workable explication of Selkirk's (1984) metric for our data.

As set out earlier, metrical phonology has provided the assumption that a syllable has a universal structure that can be defined accurately and easily as in the tree-structure template of Fig. 1. Moreover, there are restrictions in the adjacency of vowels and consonants within a syllable: The *phonotactics* of Dutch determine, for instance, that /br/ is a valid consonant couple that may occur in an onset, but that */kp/ can neither be a valid onset nor a valid coda. These phonotactics reflect both the SP (e.g. in an onset, a consonant with a high sonority cannot precede a consonant with a low sonority, as in */rbal/, or vice versa in a coda, as in */begr/), as well as more language-specific constraints relating to the articulatory restrictions of the language (e.g. in Dutch, onsets such as /tz/ or /ʃl/ and codas such as /sk/ or /z/ only occur in some rare loan words). From these within-syllable phonotactic constraints, it can be concluded that there is more statistical regularity within than across syllables.

At first sight, it may seem reasonable that a language learner can pick up the distributional probabilities of segments the language is made up. Their specific co-occurrence may then serve as the basis from which the syllable is

TABLE 1
Sonority Value Assignments to 10 Classes of Phoneme

<i>Class description</i>	<i>Phonemes</i>	<i>Sonority value</i>
a-vowels	a, a:, ɑ	1.0
e/o-vowels	e:, e, o, ɔ:, ɔ, ɔ:, eɪ, ɔu, ou	0.9
i/u-vowels	i, i:, ɪ, u:, ɪ, y, y:, u, u:, ʊ, ʊ:, œy, ə	0.8
ʌ and r-sounds	ʌ, ʌ:, ɾ	0.7
semi-vocals, laterals	l, w, h, j	0.6
nasals	m, n, ŋ, ɱ, ɲ	0.5
voiced fricatives	z, ʒ, v	0.4
voiceless fricatives	s, ʃ, f, ʧ, ʒ	0.3
voiced stop	b, d, g	0.2
voiceless stop	p, t, c, k	0.1

discovered. Moreover, the same co-occurrence learning principle could be applied to discover individual words. That is, words may be discovered because they are made up of a fixed sequence of syllables. Analogous to the discovery of intra-syllabic phonotactics, the specific co-occurrence and distributional properties of syllables may then be the bases from which words are discovered. The general idea is thus that a unit (be it syllables, words, or even idioms or syntactic structures) can be discovered from the distributional probability of its segments. In the case of syllables, it can be argued that if a large set of possible phonemes can follow a string of phonemes, then it is likely that the end of the string marks the end of a syllable; if few phonemes can follow the string, it is likely that it does not mark the end of a syllable, because intra-syllabic phonotactics constrain the number of possible phonemic continuations.

At present, there is indeed some empirical evidence that listeners use these kind of distributional cues in speech segmentation. For example, Saffran, Newport, and Aslin (1996) presented adults with synthetic speech in which the distributional probabilities between syllables were the only cues for "words" (the language consisted of six trisyllabic pseudowords like *babupu*, *bupada*, *dutaba*, etc). There were thus no prosodic cues like pauses or pitch contours in the synthesised speech sample that might have helped listeners to segment the string. Nevertheless, adults could recognise the individual words, even though these words were never presented in isolation (see also Vroomen, Tuomainen, & de Gelder, in press). Even more surprising is that eight-month-olds could compute from two minutes of speech the probability that certain syllables appear in sequence, and this allowed them to predict where one word ends and the next begins (Saffran, Aslin, & Newport, 1996). In a similar vein, Jusczyk, Luce, and Luce (1994) showed that infants are sensitive to the frequency with which certain phonetic patterns occur in the language. They used a list of monosyllabic items with frequent and infrequent phonetic patterns, and they observed that nine-month-old infants preferred to listen to the high-frequency list. These studies thus show that listeners are sensitive to co-occurrence of sounds and it suggests that listeners can learn larger elements (in this case words) from the transitional probabilities of their segments (see also Jusczyk, 1997).

In the present study, we tried to formalise a procedure that discovers the correlation existing between co-occurring segments. Moreover, we tried to explore whether this knowledge could be used as a simple segmentation device for syllables. As a first approximation, we trained a recurrent network to predict the next phoneme in a concatenated string of transcribed read-aloud words in a text. The similarity between the task of such a network and the language learner is that both "hear" words that have an internal

structure, while at the same time they do not have access to a lexicon. The issue is what kind of generalisations a network discovers from these regularities. Can this knowledge be used to acquire new words, and does it help to appreciate how the adult listener accomplishes speech segmentation?

This question was previously addressed by Cairns, Shillcock, Chater, and Levy (1997). They trained a recurrent network trained to predict phonemes on the basis of a large corpus of spoken text. One finding of Cairns et al., given the assumption that prediction errors are correlated with segmentational boundaries in the speech stream, was that the prediction error of the trained network tends to correlate more with syllable boundaries before strong syllables and content words (which is what the MSS points to as the major source of information for segmentation in English; Cutler & Norris, 1988) than with syllable boundaries before weak syllables and function words. In the present simulations, we extend the work of Cairns et al., this time using a different language (Dutch instead of English), and try to provide an in-depth analysis of the knowledge the network has acquired. The task we used was phoneme prediction because it was thought to be a task in which segmental generalisations can be picked up by the network from the input, without explicitly biasing the network via supervised learning on a segmentation task. Our primary interest lies in the automatically emerging properties of phoneme-prediction networks, rather than in their actual capability to predict the next phoneme in a string of phonemes.

EXPERIMENTS

In this section we define the phoneme-prediction task and introduce our experimental setup, namely we give a description of the data, the learning algorithm and the network topology. We then provide a description of the three analyses on the output of the trained networks, of which the results are given in the following section.

Learning the Phoneme-prediction Task

The phoneme-prediction task is defined as follows: Given a certain sequence of phonemes in connected speech, what is the identity of the next phoneme? The next paragraphs sum up our considerations on the topics of topology of the connectionist network, pattern presentation, data selection, and phoneme encoding. We furthermore provide an analysis of the task in which we express some expectations and hypotheses concerning the experiments.

Network Topology and Pattern Presentation

Various artificial neural network learning algorithms and topologies have been proposed for temporal sequence processing (for an overview, see Mozer, 1994). For instance, McClelland and Elman's (1986) TRACE model of word recognition implements a *tapped delay line* model, representing memory in the form of explicitly stored time slices. Also relevant is the work of Elman and Zipser (1988), who implement a tapped delay line model of memory in a feed-forward back-propagation network, trained on recognising and labelling phonemes when presented with a sequence of speech signal samples.

A type of topology in which copies of the complete network rather than delayed input time slices are memorised, is back-propagation through time (BPTT) (Williams & Zipser, 1990). Doutriaux and Zipser (1990) demonstrate its use in a series of simulations in which they train a BPTT network on predicting speech spectrogram time slices on the basis of previous speech spectrogram time slices. They demonstrate that sudden changes in hidden layer activity over time correlate with phonemic boundaries, that is, that the network has captured a generalisation different from the one it was trained to capture.

An interesting alternative to explicitly represented memory slices are simple recurrent networks (SRNs) (Elman, 1990), in which the representation of memory in itself is learned by the network. In its most common form, an SRN consists of a multilayer feed-forward network (usually with three layers: Input, hidden, and output layers; the input and hidden layers are fully connected, as are the hidden and output layers) with an added context layer, fully connected to the hidden layer. After each pattern presentation, the activity pattern of the hidden units is copied to the context layer, so that at the following pattern presentation, the hidden layer is confronted not only with the new input pattern activations, but also with its own contents at the time of the previous pattern presentation. Elman (1990) has demonstrated that SRNs are indeed able to perform temporal sequence processing, and, more specifically, prediction. Furthermore, Elman showed that the representations developed in the hidden layer sometimes reflect interesting facts about the task being learned.

For our simulations, we implemented a three-layer SRN. Given a training sequence of phonemes S (without any explicit syllable markers), the input layer of our network encodes phoneme S_i , and the output layer is trained to represent phoneme S_{i+1} , that is, the next phoneme in the sequence. The assumption is that as it is necessary to store a certain unknown number of phonemes in memory, the process of memory learning typical for SRN networks will learn to do so automatically. Depriving an SRN of virtually all information that would be presented when using explicit context (i.e. time

slices encoding a fixed number of previous phonemes) forces it to learn and represent just those pieces of contextual knowledge needed to solve the task. This knowledge might well be of a higher order or structure than just remembering (fading traces of) phonemes.

Data Selection

The data used in our experiments is the transcription of the first 11 pages of the novel *De Avonden* by Dutch writer Gerard Reve (1987), being read aloud. This data was introduced earlier in work on a replication of NETtalk (Sejnowski & Rosenberg 1987) using Dutch data, reported in Weijters and Hoppenbrouwers (1990). As did Weijters and Hoppenbrouwers (1990), we divided the data into a fixed training set containing the first 10 pages of the novel (4040 words, 14,955 phonemes) and a fixed test set containing the 11th page (457 words, 1616 phonemes). The training and test set thus contain sequences of phonemes (54 different phonemes occur in the data) representing connected speech without prosodic markers. This means that the individual words in the text were not pronounced in isolation, but were read aloud naturally, leading to many cross-word effects such as voicing, devoicing, and deletion of phonemes at the final and initial positions of words. For example, the end of the first sentence of the novel reads “*Frits van Egters ontwaakte*”. (“. . . Frits van Egters awoke.”). The transcription of this part of the sentence is / . . . frɪtsfanɛχtərɔntwaktə/. Voicing has occurred in the *s* of *Egters*, which in isolation would be pronounced as /ɛχtərs/; the /s/ is voiced because of the first vowel of the adjacent word *ontwaakte*, and realised as a /z/ in pronunciation. Devoicing has occurred with the *v* of *van*, pronounced /van/ in isolation; carrying over the devoiced feature of the final /s/ of the word *Frits*—/frɪts/, the *v* is realised as an /f/. An example of phoneme deletion can be found in the transcription of the same first page, on which the pronunciation of the two consecutive words *ogenblik kwam* (“a while came”) leads to a deletion of one of the two adjacent /k/s in the pronunciation /oχəmblikwam/.

We added syllable markers to the testing material by hand, for use with experimental output analyses described in a later part of this section. We calculated the occurrences of different syllabic structures (e.g. CV, CVC, etc.) in our test set. Table 2 lists the numbers and percentages of syllabic structures occurring in the testing material.

Phoneme Encoding

From Hoppenbrouwers and Hoppenbrouwers (1987) we derived a feature coding of Dutch phonemes discerning between 22 different features. We encoded these 22 features directly in both input and target patterns by setting the activations of units representing present features to 1.0, and

TABLE 2
Occurrences of Syllable Structures in the Test Set

<i>Structure</i>	<i>No. of syllables</i>	<i>% syllables</i>
CV	277	45.0
CVC	208	33.9
VC	27	4.4
CCV	25	4.1
CVCC	24	3.9
CCVC	22	3.6
V	17	2.8
CCVCC	11	1.8
VCC	2	0.3
CCCV	1	0.2

setting the activations of units representing absent features to 0.0. Taking Hoppenbrouwers and Hoppenbrouwers' phonemic features to represent an adequate and accurate feature set for Dutch, our feature encoding can be seen as an adequate coverage of Dutch phoneme space. We are thus avoiding any *direct* encoding of phonemes in input and output; when we say our SRN encodes phonemes in its input and output layer, in fact the SRN is encoding 22 articulatory features, of which none uniquely defines one single phoneme. Analogously, when we say the network predicts a phoneme, it actually predicts a set of 22 articulatory features that may not even be identical to one of the 54 phoneme encodings in our data. In those cases, we decode the output of the network to the phoneme that is the closest (i.e. has the smallest Euclidean distance) in articulatory feature space to the actual output.

Task Analysis

Phoneme prediction is a hard problem for any symbolic or subsymbolic learning system. Any such system confronted with a natural training set of concatenated phonemic strings (irrespective of the language) is faced with a multitude of possible outcomes. This high complexity of phoneme prediction stems from the fact that multiple levels of knowledge are needed to predict the next phoneme after a sequence of phonemes. This knowledge can be divided into four categories.

1. *Phonotactic knowledge*, needed to prohibit certain combinations of phonemes within syllables such as, in Dutch, */tʃ/, */ff/, and validate other combinations such as /st/, /la/, and /be/.

2. *Syllabic knowledge*, that is, knowledge about the universal structure of syllables (cf. Fig. 1), and/or about (standard) syllable sonority patterns, reflected in the sonority principle (Selkirk, 1984), which in its turn determines a major part of the phonotactics.
3. *Morphological knowledge*, needed to mark the ends and beginnings of morphemes. In Dutch, compounding of noun and verb stems is a highly productive phenomenon. Numerous cases can be found where morphological boundaries conflict with phonological principles like the maximal onset principle (MOP; consider for example the Dutch words *groe-nig* (“greenish”), where the suffix *-ig* does not interfere with the MOP, and the word *groen-achtig* (“green-like”), where the suffix *-achtig* belongs to a class of affixes that always overrules the MOP). This might be too much information to be represented as a whole by a network of the type we use. The network may however be triggered to expect at some points in the speech stream certain highly frequently occurring inflectional morphemes, such as the nominalisation affix *-ing* (realised as /ɪŋ/).
4. *Syntactic and semantic knowledge*. Although a full linguistic analysis of the speech stream would need both a syntactic and a semantic analysis, these levels of knowledge appear to be far beyond the reach of a three-layer SRN of the type we use.

In sum, it can be expected that the phoneme-prediction task will be accomplished poorly by an SRN trained on the task. This is not the point where analysis stops. An SRN trained with back-propagation learning (Rumelhart, Hinton, & Williams, 1986) always tries to minimise errors by continuously trying to find increasingly subtle ways of discovering regularities, subregularities, and exceptions, up to the point where error converges. When the error converges early and at a high level, the network may have only been able to discover very general regularities. However, these regularities may still be interesting, as they reflect inherent general properties of the data. One can posit a number of expectations on major generalisations that a network trained on phoneme prediction will make, amounting to three general hypotheses.

1. The network will discover certain phonotactic constraints generally obeyed in combinations of phonemes. This knowledge might help limiting the size of the set of possible successors when a phoneme is being presented in the input layer. The notion of phonotactics should in this case not be constrained to limitations on the adjacency of two phonemes, but to the possibility of occurrence of a phoneme after a sequence of phonemes.

2. The network will discover structural syllabic regularities. One might expect that the network will discover a regularity that is a “soft” version of the sonority principle (SP; Selkirk, 1984), that is within a syllable, sonority starts low at the onset, raises towards a peak at nucleus position, and gradually decreases during the coda until the end of the syllable, representing a hat-shaped graph. This principle bases itself more or less on the definition of sonority, of which we have adopted one in the Introduction.
3. The data used for training might contain (sequences of) words occurring more than once in the text; furthermore, it may contain certain frequently occurring affixes in morphologically complex words. Both features of the text may prompt the network to pick up on these extra-syllabic lexical and morphological regularities.

The discovery of certain phonotactics and the SP, mentioned in hypotheses 1 and 2, would help minimise prediction error if the syllabic structures in the data were of constant size and type (e.g. CV). Yet, in the case of our data set containing natural running speech, additional difficulties for predicting the next phoneme are the irregularly situated boundaries between syllables, morphemes, and especially words. At these boundaries, practically any phoneme can be expected. Consequently, when an SRN has discovered phonotactics and the SP, it will be able successfully to apply these generalisations only *within* the smallest phonotactically coherent phonemic groups, namely syllables. Predicting phonemes within a syllable should be increasingly easy going from the beginning of the syllable to the end, since the remembered left context becomes more coherent as more phonemes belonging to the same syllable are presented and remembered. This can be visualised by a sawtooth-like error graph, peaking at syllable onsets and gradually decreasing, until the next syllable begins and a new peak emerges. Error in this respect should be defined in terms of the distance between the output phoneme and the desired phoneme, for example, Euclidean distance in phonemic-feature space. Ideally, the error graph computed on the basis of the SRN output would consist of sequences of these sawtooths.

This error graph can then be used as an input source for a syllable boundary detector. The accuracy of this boundary detector can be tested by counting coinciding error peaks and real syllable boundaries. When this score turns out to be significantly larger than an averaged random baseline score, the network proves its usefulness on a task different from the one it was trained on. When the sonority principle is discovered by the SRN, a similar syllable detection strategy can be devised. A sonority graph can be computed on the basis of the identities of the predicted phonemes, that ideally would consist of sequences of alternating ascending and descending graphs. Wherever a descending line ends and an ascending line begins, a

syllable boundary can be expected. That syllable boundary might very well be one position to the left or to the right, since bottom sonority might lie on (a) the last consonant of the coda of a previous syllable, (b) the first consonant of the new syllable, or (c) the second consonant of the new syllable in case of an exceptional (extrametrical) /s/ in the onset. Figure 2 displays three two-syllabic example words demonstrating these three possibilities.

In the word displayed in Fig. 2(a), /ratsal/ (*raadzaal*, “council hall”), it can be seen that the syllable boundary between /t/ and /s/ follows the bottom sonority of the /t/. In Fig. 2(b), /lertak/ (*leertaak*, “learning task”), a syllable boundary occurs immediately before the /t/ with the bottom sonority. Finally, the word displayed in Fig. 2(c) /narstə/ (*naarste*, “nastiest”), contains an extrametrical /s/ in the final syllable /stə/.

Experimental Setup and Methods

In this subsection we provide details on the SRN learning parameters and topology used in our experiments, as well as a description of the analyses performed on the experimental outcomes.

Learning Parameters and Topology

On the basis of the fixed training and testing material, we performed 10 experiments with different weight initialisations (weights were set randomly at values between -1.0 and 1.0). In each experiment, the SRN was trained with a learning rate of 0.1 , a momentum of 0.4 , 22 input units, 60 hidden units,² 22 output units, a convergence threshold of 0.0001 with a convergence patience of 2 (i.e. training was stopped when the mean squared error of the SRN decreased by 0.0001 or less during two training cycles, where one training cycle equals the presentation of the full training set to the SRN), and an update tolerance of 0.2 (i.e. weights of connections from an output node of which the activation is within 0.2 of its target value are not updated in back-propagation).

Figure 3 presents a graphical representation of the network topology used. In this example, where the phoneme sequence /bɑl/ (*bal*, “ball”) is being processed, the network correctly predicts an /l/ after the presentation of /ɑ/, while somehow remembering the previous presentation of /b/ in its context layer.

²A fixed hidden layer of 60 hidden units was chosen on the basis of a number of pilot experiments using only the training material and monitoring the convergence of mean squared error.

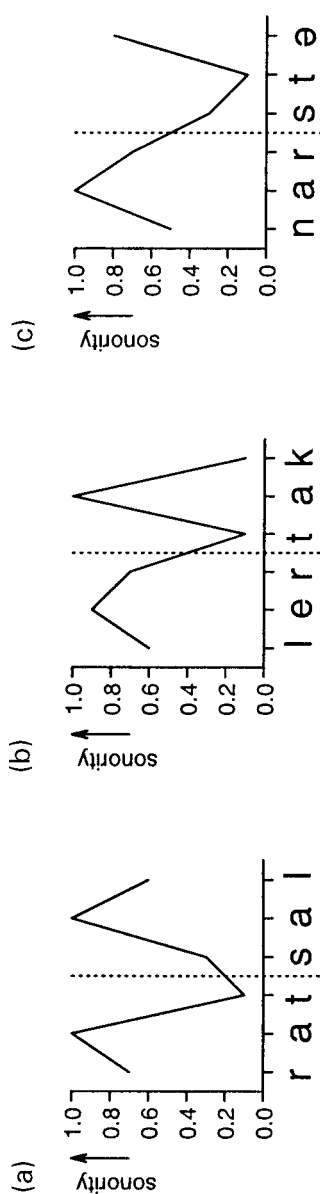


FIG. 2. Three examples of two-syllabic words with a syllable boundary (a) that occurs immediately after the phoneme with the lowest sonority; (b) that occurs immediately before the phoneme with the lowest sonority; and (c) that, because of an extrametrical /s/, occurs two positions before the phoneme with the lowest sonority.

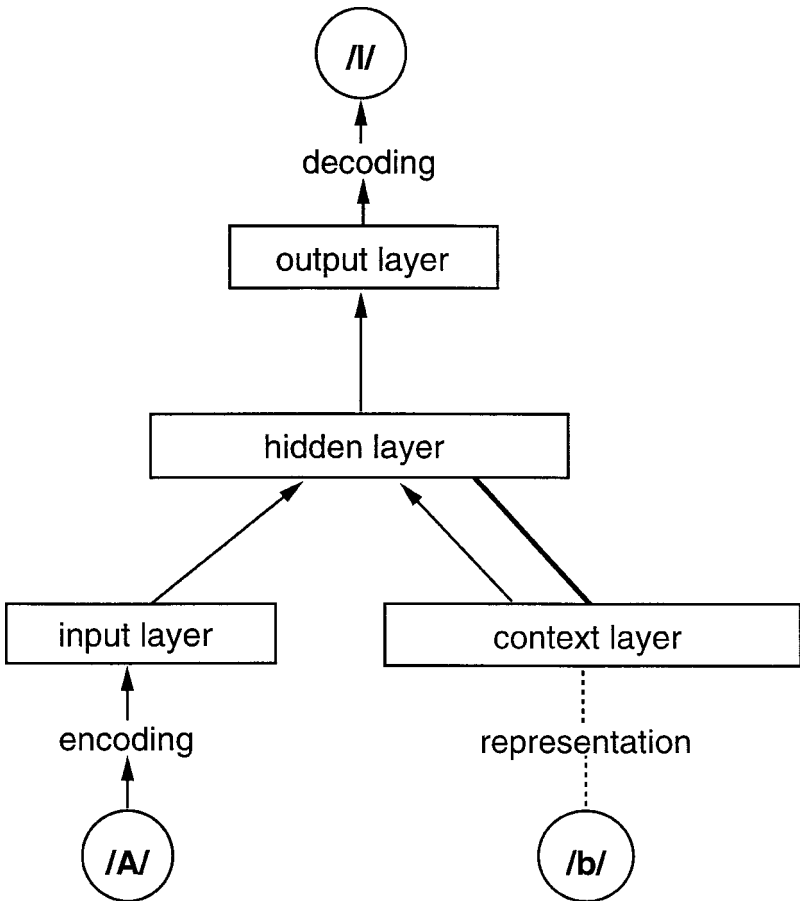


FIG. 3. Network topology used in experiments. The network has three layers plus an extra context layer. Arrows between layers represent standard back-propagation connections (layers are fully connected); boldface type line represents a full one-to-one copy connection between the hidden and the context layer.

Analyses

In order to be able to test the first and second hypothesis stated previously we performed three analyses on trained networks.

1. *Prediction-error analysis*: Decoding the output of a trained network into a phoneme and compare it with the target phoneme, to see how often phonemes are predicted correctly relative to their position in the syllable.

2. *Sonority-graph analysis*: Investigating the compatibility of the network's output with the SP, by decoding the phonemes produced by the phoneme prediction analysis into sonority values (between 0.1 and 1.0), measuring mean sonority values for all syllable lengths and positions, and calculating the number of correctly produced sonority patterns using the syllable boundaries of the target sequence.
3. *Syllable-boundary analysis*: Investigating the relation between cues in the network output and syllable boundaries in the material. An analysis is made of the occurrence between syllable boundaries and (a) phonemic-feature distances between output and target phonemes, and (b) transition points from decreasing to increasing sonority values in the phoneme sequence produced by the network.

RESULTS

Prediction-error Analysis

The network output is decoded into a phoneme each time a pattern is fed forward through the network. Decoding is done by computing the Euclidean distance between the output activation values and the phonemic feature codes of all 54 phonemes, and taking the phoneme associated to the code with the smallest Euclidean distance as the output of the network. When this phoneme is identical to the target phoneme, it is counted as correctly predicted.

An illustrative baseline score for predicting the identity of a phoneme without knowledge of its left context, is 11.9%, namely the percentage occurrence of the most frequently occurring phoneme in the test set, /ə/. The average prediction score over 10 experiments on the test set was 22.5% (standard deviation: 0.6): as expected, prediction accuracy is low, but is clearly more accurate than the result of just guessing the most frequent phoneme; the score surplus does indicate some success in generalisation.

Counting the numbers of correctly predicted phonemes, and sorting these numbers by the syllable positions they occur in, would give a rough image of the distribution of prediction errors within the syllable. The dissatisfactory aspect of this crude prediction, however, is that it would not be able to express the fact that, for example, an output of /a/ would be less incorrect in the case of the target value /a/ than it would be in the case of the target value /p/. We therefore introduce a prediction-error function sensitive to the fact that some phonemes are more alike. First, a definition of this prediction error is needed. The prediction error of a single network output (i.e. the prediction of one phoneme) is the Euclidean distance in phonemic-feature space between the feature encoding of the output phoneme and the feature

encoding of the target (desired) phoneme (a positive integer value).³ Thus, the prediction error simply expresses the number of articulatory features the prediction is off the target. For example, if the network predicts a /a/ while the target phoneme is /a/, the prediction error is 1, as /a/ differs in only one articulatory feature from /a/ (namely, the back feature, expressing the place of articulation in the mouth). In case of the network predicting a /a/ when the target phoneme would be /p/, the prediction error would be 10, as /a/ and /p/ differ in 10 articulatory features.

We computed over all experiments the average phonemic-feature distance per syllable position; Fig. 4 displays these values for the overall values averaged over all syllable structures, and for the five most frequently occurring syllabic structures. The figure displays an overall decreasing phonemic-feature distance from the beginning of the syllable to the end of

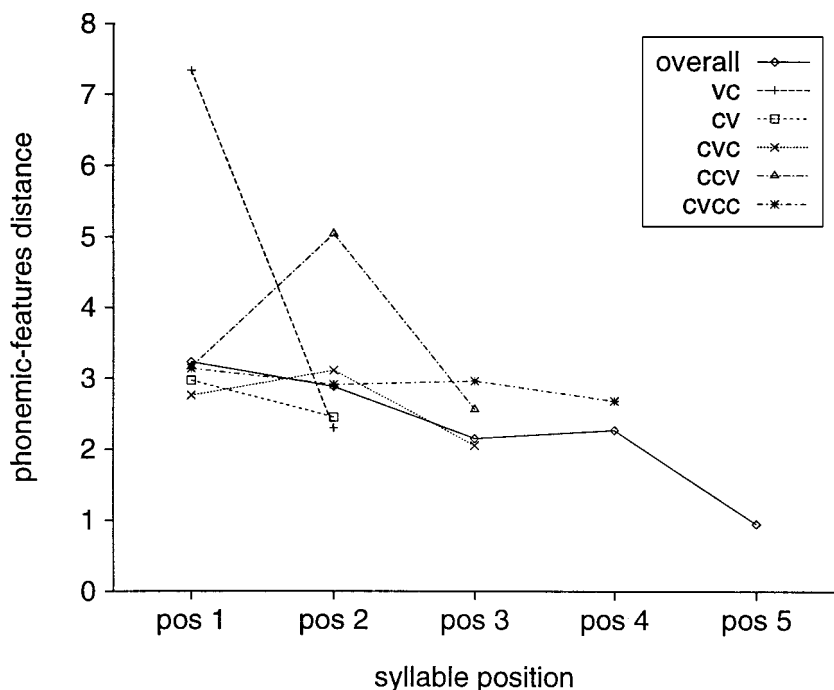


FIG. 4. Percentage prediction-error values (phonemic-feature distances) of predicted phonemes of the five most frequent syllable structures and an overall average, per syllable position.

³Elman (1990) implements a similar prediction-error analysis in his experiments with the application of SRNs to letter prediction in sequences of words.

the syllable, with a minor increase on the fourth syllable position (due to the fact that on this position mainly consonants occur; as more features are devoted to distinguishing between consonants than between vowels, consonants have on the average a larger distance to each other). More generally, the results in Fig. 4 show that the prediction of the final phoneme in a syllable is always more accurate than the prediction of the first phoneme.

Figure 4 also displays the fact that the largest prediction errors are made with the V of the VC structure. As the prediction error is about 7.5 (i.e. there are on average about seven–eight differing articulatory features between the target phoneme and the predicted phoneme, which is in fact equal to the average distance between vowels and consonants in the Hoppenbrouwers & Hoppenbrouwers, 1987 encoding), the network on the average expects a C. This is understandable given the fact that within the training set only about 7.5% of the syllables start with a V. After the actual presentation of the V of a VC structure, the network is able to produce a much better prediction, which is mostly a C; the large difference for the V and C for VC structures in Fig. 4 is a reflection of this fact. A second deviation from the overall decreasing trend in Fig. 4 is the relatively large phonemic-feature distance of the prediction of the second C in CCV structures. Apparently, the network is mostly expecting a phoneme quite different from a C there, that is, a V, as it may have generalised to expect a V after a C, which is indeed the majority case in the data: In 61.5% of all cases, a C is followed by a V, and in 38.5% by a second C. A third deviation in Fig. 4 is the fact that the V of CVC structures is predicted with a lower accuracy than the seemingly similar V of CV structures. Although the difference is less than one articulatory feature, it can be explained by the fact that more different Vs can occur in the closed syllable structure CVC (i.e. both long and short vowels) than in the open syllable structure CV (i.e. mainly long vowels and /ə/). Having to choose between more alternatives in predicting the V of a CVC structure may thus induce slightly larger prediction errors in terms of articulatory features as compared to predicting the V of a CV structure.

In metrical phonological theory, the difference between the frequently occurring “majority” cases, such as the CV structure in syllables, and the less frequent cases, such as VC and CCV structures, is often expressed as a difference between *unmarked* (i.e. general) and *marked* (i.e. exceptional) cases. This difference serves to trigger the application of general phonological rules to the unmarked cases, and the application of specific (exceptional) phonological rules to marked cases. Markedness of a segment is thus a means to select the proper rules that should apply to that segment (Calabrese, 1995). It is the task of the human listener/speaker to learn to recognise markedness, which in general is more difficult to do than to learn the underlying principles for the unmarked cases. Although our network is in no way explicitly trained to recognise markedness, it does display a

tendency to expect the unmarked case (i.e. the CV syllable structure), and to have trouble recognising marked cases (such as the VC and CCV structures). It should be noted that this superficial correlation is nothing more than an emerging property of the application of the total of all distributional generalisations the network has learned.

Focusing solely on syllable structures does not give right to the fact that the network actually learns to predict successfully a limited number of phoneme strings occurring frequently in the training material (this fact relates to the third hypothesis mentioned previously, stating that the network will learn frequently occurring words or morphemes in the text). Close analysis of the predicted output shows that the network is able to predict with high accuracy the final part of the first name of the main character in the novel, /frɪts/ (*Frits*). Generally, after /f/ and /r/ are presented (and mostly mispredicted), the network consistently produces /t/, /t/, and /s/ as the three next phonemes. The same applies to the words /mudər/ (*moeder*, “mother”) and /va:dər/ (*vader*, “father”), for which the network is consistently able to predict /dər/ after having seen /m/ and /u/, or /v/ and /a:/, respectively. No similar effects were found relating to high-frequency affixes such as /ɪŋ/ *-ing* or /tə/ *-te*, as they are simply too short, and their preceding phonemes and phoneme strings too diverse to trigger uniquely correct predictions. Yet, from the three examples /frɪts/, /mudər/, and /va:dər/, relating to the three most frequently occurring characters in the novel *De Avonden*, it can be seen that the network can learn a limited number of phoneme sequences relating to high-frequency lexical items.

Sonority-graph Analysis

As set out earlier in the Introduction, the sonority principle (SP; Selkirk, 1984) states that, within a syllable, sonority starts low at the onset, increases towards a peak value at the nucleus position, and gradually decreases along the coda until the end of the syllable. We computed the average sonority values of all predicted phonemes and sorted them according to their positions in the syllables of the corresponding target phoneme sequence. Figure 5 displays these average sonority values for the five most frequently occurring syllable structures. The graph in Fig. 5 is centred around the nucleus of each syllable. If the network has grasped the sonority principle, one would expect to see a peak at the nucleus, and decreasing sonority values going further right or left into the syllable. This general pattern is indeed present in Fig. 5 for each of the five syllable structures, albeit weakly in the case of VC-structures. The relatively low predictability of the V of VC syllables is expressed in Fig. 5 by an average sonority value of 0.44, indicating that in the network usually expects a consonant at that point (cf. Fig. 4 for an analogous effect in phonemic feature distance at the same position).

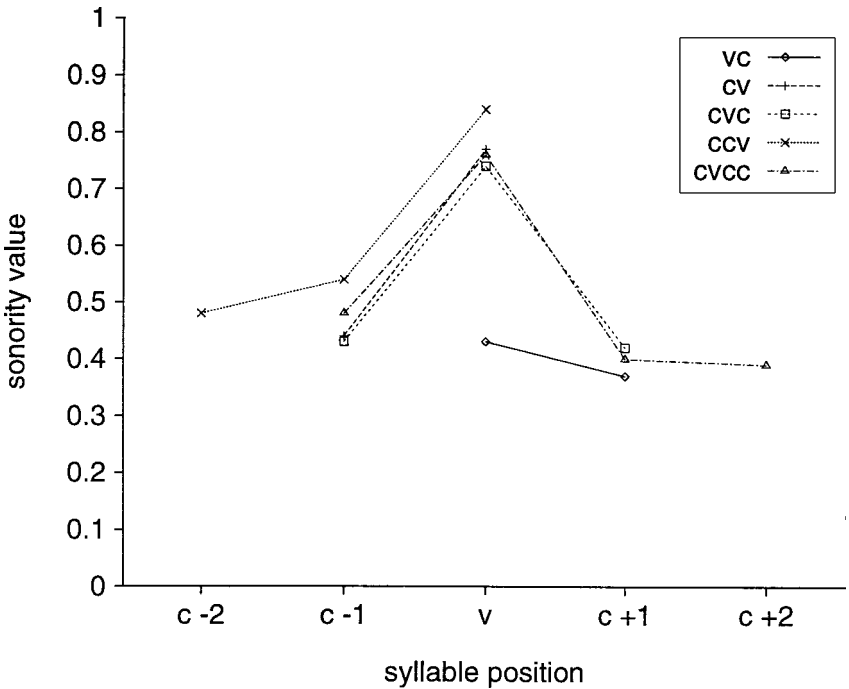


FIG. 5. Average sonority values of predicted phonemes, per syllable position for each of the five most frequently occurring syllable structures. Graphs are centred around the nucleus.

The results displayed in Fig. 5 for the CV, CVC, and VC syllable structures leave the interpretation open that the network is constantly predicting a CVCVCV... sequence. This would be normal since CV is the most frequently occurring syllable structure (45.0% of the syllables in the data have a CV structure); the most probable phoneme following a C is a V, and vice versa. By predicting a CVCVCV... sequence, CV and CVC syllables could be correctly predicted as regards the sonority of their phonemes. However, the average sonority values plotted for the CCV and CVCC syllables show that the network is able, on average, to predict that the second phoneme of the CC clusters in both the CCV and the CVCC structures is indeed a consonant, rather than a vowel. Both sonority values reflect that the network is expecting a consonant with sonority 0.4 (on average). It can therefore be hypothesised that the network learns to be sensitive to context to a degree that enables it to be more subtle than predicting only a CVCVCV... sequence. On inspection, it turns out that there are statistical

clues in the left-context of the second C of CC clusters both in CCV and in CVCC syllables for expecting a consonant, that the network may have learned. For the case of CCV syllables, the first consonant has an average sonority of 0.20 in our data. Chances are that after a low-sonority consonant (of sonority 0.3 or lower) in syllable-initial position, a higher-sonority consonant (of sonority 0.4 to 0.6) or a low-sonority vowel (of sonority 0.7 or 0.8) follows, rather than a high-sonority vowel (sonority 0.9 or 1.0). This is the case in 76% of all syllables starting with a consonant with sonority value 0.3 or lower. The network appears to have picked up this generality by expecting, on average, a high-sonority consonant rather than a vowel as the second phoneme in CCV syllables. Although this suggests that the second phoneme of CCV syllables is predicted rather accurately, the prediction accuracy of this phoneme displayed in Fig. 4 shows that the actual predicted phoneme is, on average, rather different from the target phoneme. Sonority appears to be predicted better than articulatory features here.

For the case of CVCC structures, both the V and the C before the second C in the CC clusters provide clues to expect a low-sonority consonant. First, the V is generally short in CVCC syllables: 18 of the 24 CVCC syllables in our data have short vowels. Second, the first C of the CC cluster in CVCC syllables has an average sonority value of 0.54. Moreover, the second C has an average sonority value of 0.17 (15 CVCC syllables have /t/ in final position). Although the data contains many exceptions, the rule appears to be that a low-sonority consonant can be expected after a short vowel and a high-sonority consonant (this is the case in 18 out of the 24 CVCC syllables). By predicting, on average, a consonant with sonority 0.39 as the final phoneme of CVCC syllables, the network shows that it is indeed able to use the context to predict consonants more often than vowels at this position, thus overruling the general bias towards predicting a CVCVCV... sequence.

Altogether, Fig. 5 provides indications of the fact that the output of the network might reflect the SP. As a direct test of the compatibility of the output of the network with the SP, we compared the sonorities of the sequence of phonemes produced by the network to the syllables of the target phoneme sequence. For each syllable in the target sequence, we checked (a) whether the corresponding sonority pattern of the predicted phoneme output peaked at the position of the nucleus, and (b) whether it monotonically decreased further away left and right from the nucleus, that is, whether that syllable abided by the SP. We found that the percentage of sonority patterns produced by the network abiding by the SP in this way was 92.6% (averaged over the 10 experiments). For comparison, the target

phoneme sequence (the test set) contains 96.7% SP-abiding syllables.⁴ This remarkably high percentage of SP-abiding sonority sequences in the network output indicates that the generalisations the network has made during training reflect the SP to a rather high degree.

As an illustration, Fig. 6 displays an excerpt from the test set of which the target string reads /zəʎɪŋna:rdəkœykəhuɣa:təθi:r/ (*Ze ging naar de keuken. "Hoe gaat het hier . . ."* She went to the kitchen. "How are things here . . ."). The characters displayed in Fig. 6 representing phonemes are directly taken from our ASCII-encoding of the International Phonetic Alphabet. It can be seen that only a few phonemes are predicted correctly (8 out of 24), most of them at the end of syllables (of which the boundaries are denoted by vertical dotted lines). More important is the graph part of Fig. 6, showing both the sonority values of the target string of phonemes (dashed line) and the sonority values of the predicted phoneme string (solid line): It can be seen that for all syllables in the target string, *both* lines reflect the SP; furthermore, the network is very well capable of predicting VC structures. On two occasions in this example string, one consonant follows another consonant in the target string; in the first case, the network incorrectly expects a V (i.e. a /ə/). However, according to the criterion mentioned previously with which we compute the amount of syllables in the string of predicted phonemes that abide by the SP, all syllables in the string of predicted phonemes follow the SP.

Syllable-boundary Analysis

We have shown that: (a) SRNs trained to predict phonemes in phoneme sequences display the ability to predict phonemes at the end of a syllable more accurately than phonemes at the beginning of a syllable, and (b) trained SRNs produce sequences of predicted phonemes that abide by the SP to a high degree. Both phenomena could contribute to finding segment cues, for example, syllable boundaries, in the SRN's output. Before reporting on correlations found between these phenomena and syllable boundaries, it is important to establish a baseline criterion for success in finding such correlations. We measure this success in terms of the accuracy of deciding whether a phoneme is the first phoneme of a syllable. Four events can occur with such decisions: (1) The decision is correctly YES, that is, a hit; (2) the decision is incorrectly NO, that is, a miss; (3) the decision is incorrectly YES, that is, a false alarm; or (4) the decision is correctly NO, that is, a correct rejection. The percentage of correct syllable boundary

⁴The 3.3% defective cases mainly stem from the fact that in Dutch the onset /st/ and the coda /ts/ are accepted, whereas the sonority of /s/ is usually taken to be higher than that of /t/. As set out earlier, linguistic theories of metrical phonology usually regard the /s/ as being extrametrical to account for this irregularity (cf. Halle & Vergnaud, 1980).

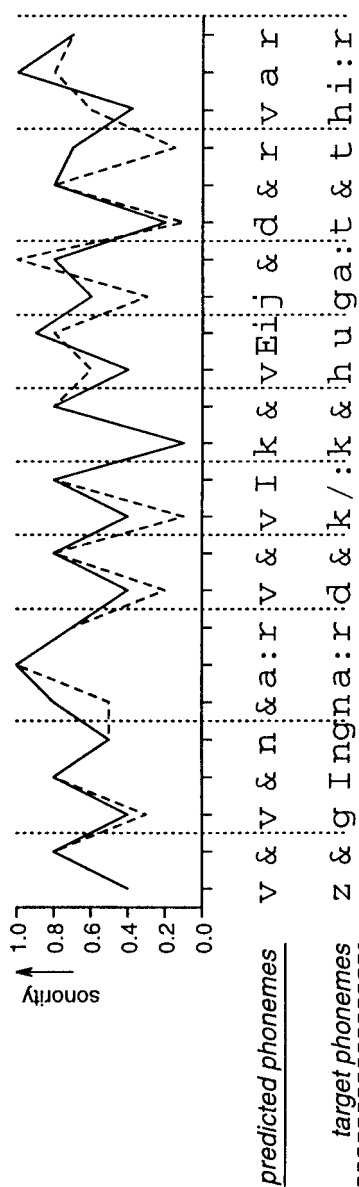


FIG. 6. Examples of a string of target phonemes extracted from the test set (bottom row of characters), the string of phonemes predicted by the network (top row of characters), and the sonority values of both strings plotted in the top graph area. The solid line represents the sonority graph of the string of predicted phonemes; the dashed line represents the sonority graph of the string of target phonemes. Syllable boundaries are denoted by vertical dotted lines.

decisions is calculated by taking the portion of the sum of hits and correct rejections in the total number of decisions. Always predicting the most frequent outcome, NO, results in a baseline accuracy of 60.9% correct decisions. Any accuracy resulting from basing syllable boundary decisions on the basis of the SRN's output should be significantly higher than this baseline accuracy to be considered relevant.

Counting the occurrence between prediction-error peaks and syllable boundaries quickly showed that prediction error (i.e. Euclidean phonemic-feature-space distance) does not provide sensible clues for the occurrence of syllable boundaries. Counting an error difference between the prediction error of a phoneme and the previous phoneme of three or larger as an error peak, we obtained an optimal decision accuracy of 62.0%, hardly better than the baseline score. Apparently, there is too much noise in the phonemic output of the network for it to be used as a source for syllable boundary cues.

Sonority Changes and Syllable Boundaries

When the sonority of the phoneme predicted by the SRN is lower than that of the preceding phoneme *and* the next phonemes, chances are that a syllable boundary occurs immediately before or after that phoneme. It was shown earlier that the SRNs were generally able to grasp the sequential aspects of the SP, hence some success might be expected here. We concentrated on analysing the co-occurrence of syllable boundaries in the output phoneme sequence, and phonemes in the network output at position P with lower sonority values than their immediately preceding and following adjacent phonemes at positions $P - 1$ and $P + 1$. This strategy led to a syllable boundary decision accuracy of 79.9%, significantly better than the baseline accuracy [$t(10) = 37.4, P > 0.001$]. It should be noted that the same strategy applied to the *target* phoneme sequence leads to a decision accuracy of 82.1%.

We applied a second, less stringent method to our data, which counts a decision as correct when a syllable boundary is at the same position as the output phoneme with lower sonority than its neighbours, *or* one position immediately following this position. This method gives right to the fact that consonants with a low sonority may well be the last consonant of a coda, therefore marking the position immediately *before* the syllable boundary (cf. Fig. 2 for examples). This method renders a decision accuracy on the basis of the output of the networks of 91.1%, significantly better than the baseline output, $t(10) = 77.8, P > 0.001$. Although this score represents a flattering decision accuracy, it indicates that the SRNs have grasped the hat-shaped behaviour of sonority along phoneme sequences.

Altogether, these results indicate that *some* information concerning syllable boundaries is present in the sonorities of predicted phonemes. This

information is often incorrect, but gives significantly more clues than a baseline guess about the presence and positions of syllable boundaries.

CONCLUSION

The present study investigated the output of SRNs which, although explicitly trained on a different task, are able to extract basic generalisations that hold over sequences of phonemes in connected speech. We showed that one of the most apparent generalisations of the trained SRNs is the fact that sonority varies in a hat-shaped way, marking syllable boundaries. Although the hats predicted by the networks did not match all syllable boundaries in the target phoneme sequence, they do abide by the SP in 92.6% of the syllables. The network thus succeeded in discovering some important structural aspects of unsegmented speech. It should be noted that this “discovery” was neither part of the task the network was trained on, nor an explicit content of the network output. Rather, we have used analyses on top of the network to generalise over the output of the network after training. When we claim that the network has discovered the SP, we mean that analyses show that the output of the network abides by the SP to a high degree, while the network’s SP-abiding behaviour actually emerges from the application of the total of all distributional generalisations the network has learned.

Some conclusions can be drawn regarding the bootstrap problem mentioned in the Introduction. We wanted to examine whether a recurrent prediction network was able to give segmental cues while processing unsegmented speech. With an approach comparable to the approaches Cairns et al. (1997), Doutriaux & Zipser (1990), Elman (1990), we looked for indications of generalisations made by the network that were *indirectly* related to the task of prediction. As said, one such generalisation we found was the sonority principle. Although the cues extracted from the network regarding syllabic structure were far from perfect, the discovery of the sonority principle is a major step towards the discovery of the syllable. This knowledge, on its turn, may help the human listener to segment speech into words.

However, one of the critical aspect of this claim concerns the relation between the task the model is performing and the task the human listener is doing. Is it the case that the child is predicting phonemes and then discovers, by accident, the “real” principles of language? There is some evidence from the phoneme monitoring literature that adults can indeed predict the upcoming phoneme of a word (e.g. Marslen-Wilson, 1984). However, the phoneme-monitoring task can hardly serve as an analogy for listening to speech, let alone for the task the newborn is confronted with when listening to an unknown language. So, in fact, it seems unlikely that the child is doing

anything that resembles phoneme prediction. On the other hand, it seems likely that the task the network is doing is not that critical, because the discovery of the sonority principle has also been found with different networks in different tasks environments. For example, Corina (1994) used a SRN for root duplication within an artificial language. He observed that, although the network performed only moderately on the task at hand, it had nevertheless discovered knowledge about syllabic well-formedness. It was also argued that the network might have extracted sonority information from the input, even though it was not explicitly coded in the input. Given that a similar type of information is discovered in a different task, it seems that knowledge about syllabic well-formedness, and more in general, the induction of language regularities or principles, is not crucially dependent on the task the network is performing. There is therefore no need to maintain that the child is doing something like phoneme prediction while listening to speech. Rather, the conjecture is that even though the task of the network is different from that of the child, both may discover the same co-occurrence principles in the speech signal.

The analogy between the simulations presented here and the child learning the language is thus a loose one. Moreover, the newborn has other cues to speech segmentation than distributional probabilities of phonemes. Strong as opposed to weak syllables are one possibility (e.g. Cutler & Norris, 1988; Vroomen & De Gelder, 1995, 1997), vowel identity in languages with vowel harmony are another (Suomi, McQueen, & Cutler, 1997; Vroomen, Tuomainen, & de Gelder, in press), but also phonetic cues like word-final vowel lengthening (Klatt, 1975), or stress in fixed-stress languages, may serve as additional cues to word segmentation (Vroomen, Tuomainen, & de Gelder, in press). In general, it seems likely that language learners take advantage of the many cues they can discover in the language input, because eventually prosodic and distributional cues will converge on the same word boundaries.

REFERENCES

- Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus based approach to speech segmentation. *Cognitive Psychology*, *33*, 111–153.
- Calabrese, A. (1995). A constraint-based theory of phonological markedness and simplification procedures. *Linguistic Inquiry*, *26*, 373–463.
- Corina, D.P. (1994). The induction of prosodic constraints: Implications for phonological theory and mental representations. In R. Corrigan, G. Iverson, & S. Lima (Eds), *The reality of linguistic rules*. Philadelphia, PA: John Benjamin.
- Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 113–121.
- Doutriaux, A., & Zipser, D. (1990). Unsupervised discovery of speech segments using recurrent networks. In D.S. Touretzky, J.L. Elman, T.J. Sejnowski, & G.E. Hinton (Eds), *Connectionist models: Proceedings of the 1990 summer school*. San Mateo, CA: Morgan Kaufmann.

- Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.
- Elman, J.L., & Zipser, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America*, *83*, 1615–1625.
- Frauenfelder, U., & Peeters, G. (1990). Lexical segmentation in TRACE: An exercise in simulation. In G.T.M. Altmann (Ed.), *Cognitive models of speech processing*. Cambridge, MA: Bradford Books.
- Halle, M., & Vergnaud, J.-R. (1980). Three-dimensional phonology. *Journal of Linguistic Research*, *1*, 83–105.
- Hoppenbrouwers, C., & Hoppenbrouwers, G. (1987). De featurefrequentiemethode en de classificatie van Nederlandse dialecten. *Tabu*, *18*, 51–92.
- Jusczyk, P.W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.
- Jusczyk, P.W., Luce, P.A., & Luce, J.C. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645.
- Kiparski, P. (1979). Metrical structure assignment is cyclic. *Linguistic Inquiry*, *10*, 411–441.
- Klatt, D.H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *3*, 129–140.
- Marslen-Wilson, W.D. (1984). Function and process in spoken word recognition. In H. Bouma & D.G. Bouwhuis (Eds), *Attention and performance: Vol. X. Control of language processes*. Hove, UK: Lawrence Erlbaum Associates Ltd.
- Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. *Cognition*, *25*, 71–102.
- McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.
- Mehler, J., Dommergues, J., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behaviour*, *20*, 298–305.
- Mozer, M.C. (1994). Neural net architectures for temporal sequence processing. In A. Weigend & N. Gerschenfeld (Eds), *Time series prediction: Forecasting the future and understanding the past*. Reading, MA: Addison-Wesley.
- Reve, G. (1987). *De avonden: een winterverhaal*, (34th edn.). Amsterdam: Bezige Bij.
- Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations* (pp. 318–364). Cambridge, MA: MIT Press.
- Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.
- Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.
- Sejnowski, T.J., & Rosenberg, C.R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, *1*, 145–168.
- Selkirk, E.O. (1984). On the major class features and syllable theory. In M. Aronoff & R.T. Oehrle (Eds), *Language sound structure* (pp. 107–136). Cambridge MA: MIT Press.
- Suomi, K., McQueen, J.M., & Cutler, A. (1997). Vowel harmony and speech segmentation in Finnish. *Journal of Memory and Language*, *36*, 422–444.
- Trommelen, M. (1981). Dutch diminutive formation as a rime-bound process. *The Linguistic Review*, *1*, 345–368.
- Vroomen, J., & de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 98–108.
- Vroomen, J., & de Gelder, B. (1997). Activation of embedded words in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 710–720.

- Vroomen, J., Tuomainen, J., & de Gelder, B. (in press). The roles of word stress and vowel harmony in speech segmentation. *Journal of Memory and Language*.
- Weijters, A., & Hoppenbrouwers, G. (1990). Netspraak: een neurale netwerk voor grafeem-foneem-omzetting. *Tabu*, 20, 3–28.
- Williams, R.J., & Zipser, D. (1990). *Gradient-based learning algorithms for recurrent connectionist networks*. Tech. Report No. NU-CCS-90-9. Boston, MA: Northeastern University.