THE NETWORK LEXICON:
A NOVEL APPLICATION OF PHONOLOGICAL KNOWLEDGE IN ASR

Mark HUCKVALE

Abstract

This paper describes an architecture for Automatic Speech Recognition that uses phonological knowledge in a novel way. The architecture, a descendant of the connectionist TRACE model of speech perception consists of a network of words activated from below by a phonetic component and uses phonology to specify links between words so that words having similar phonological description share activation. This leads to improved discrimination performance between similar words, and the phonological representation arising from an analysis of an utterance can be seen to be related to how the whole lexicon responds to the utterance.

The paper outlines the word and phonological layers, proposes a phonetic component based on pattern recognition principles, and discusses how the junction of the two relates to existing speech recognition principles.

1. Introduction

The current application of phonological knowledge in Automatic Speech Recognition is mainly limited to two areas:

a) as a specification for an intermediate representation for the utterance to be recognised, between the acoustic signal and the word sequence (or lattice). For example Kohonen et al (1987), Waibel et al (1989).

b) as a specification for a set of acoustic models required to analyse an utterance into a sequence of phonetic events. For example, Lee et al (1989), Levinson (1985).

Both have these applications have faults; since a) speech signals are not easy to segment and label, so traditional linear segmental accounts of the utterance over-simplify the acoustic structure; and since b) the acoustic form of phonetic events is highly variable as a function of speaker, environment, context and occasion. Both of these faults are discussed in more detail in Huckvale (1990).

In this paper I should like to outline an architecture for ASR which uses phonological knowledge in a novel way. Instead of interposing phonological representations between signals and words, the architecture places a phonological layer above the word level, so that a phonological representation emerges as a consequence of how the lexicon reacts to an unknown utterance, with the words themselves being activated directly from a phonetic component operating on the signal. In section 2 I shall describe the word and phonological layers, in section 3 the phonetic component, and in section 4, the connection between the two.

The architecture is a direct descendant of the lexical layer in the TRACE model of speech perception (McClelland and Elman, 1986), and some familiarity with TRACE is assumed.

2. The Word and Phonological Layers

Word Units. We start with a single layer of units, each representing a word, and each fed from below by a phonetic component (described in sections 3 and 4) which activates words with a probability of their presence in the input signal at the current instant. Since only one word is required for each instant, the units within this layer are interconnected with inhibitory links so that they compete with one another, and the degree of competition relates to the extent to which they explain similar regions of the signal.

A recognised word sequence in such a lexicon is the sequence of maximum activations of the word units. The sequence is either in time - following activations as they develop in a single layer, or in space - following activations through a sequence of layers. The TRACE model replicated the lexicon in space, and separated network 'time' - the computational time required to communicate activations, from speech signal 'time' - the development of a linguistic sequence. Clearly we would like to reconcile these two 'times' in a future, more cognitive, system.

Such a system already has interesting properties, as TRACE has shown. The competition between words to explain a segment of the signal ensures that explanations of the signal will consist of words which cover the signal and which do not overlap to any large degree.

However, since each word is fed independently from below by the phonetic component, the word layer alone will not have very good discriminating power for similar words. Thus phonetic evidence arising from an analysis of the word 'bin' might activate words 'bin' and 'pin' to a pretty equal amount. Final choice between them has to be made on the overall word difference, which does not take into account that the two words have a large number of phonological similarities. Thus 'pin' might be chosen, erroneously, if its second half happens to be more similar to the input than the second half of 'bin'. Speech recognition systems use sub-word units to circumvent this problem: to structure the acoustic similarities between words and to concentrate discriminations onto the elements of the words which are known to be used to make phonological distinctions.

Phonological Units. The problem with traditional systems is that the phonological model is used to structure the acoustic/phonetic evidence, rather than to demonstrate choice in the lexicon. In the network lexicon we add phonological analysis as additional layers of units on top of the words which tie together words having similar phonological prescription. The choice of units will depend on the most favoured phonological theory, but we will assume there

will be units representing syllabic structure, stress patterning of words, consonant clusters, bilabiality of syllable onset, nasality of syllable coda, as well as many others. These units have positive bidirectional connections with the word units. Thus if 'bin' is activated, then some of that activation is shared (in different amounts) with all other syllables ending in '-in', all other monosyllabic words, all syllables containing a front vowel, all syllables with a single initial consonant, etc. The spread of activation is based on our preferred phonological theory, not on measurements of acoustic similarity. See Figure 1.

The phonological units enhance discrimination power in the following elegant way. Since the phonetic evidence activates all words to some degree, the phonological interpretation of the phonetic evidence is expressed in the pattern of activation of the words. If some phonetic evidence is for syllable final nasality, then many words containing syllable final nasals will be activated from below, which in turn will activate phonological units representing syllable final nasality. Similar phonological activations will result from the activations of words containing syllables beginning with /p/ and /b/. The degree to which 'bin' is chosen over 'pin' therefore, is a function of the activation feeding from 'voiced syllable initial' and 'voiceless syllable initial' phonological units. (You can think of it that all words ending in '-in' have roughly equal activation through sharing of phonological properties, and hence recognition is focussed on the phonological dissimilarities of the two words). The activations of these units in turn are a function of how the whole lexicon has reacted to the phonetic evidence. Thus just as phonological analysis is based on the structure of the lexicon, the emergent phonological representation is based on the response of the lexicon to the phonetic evidence. The sequence is: phonetic-evidence to word activations to phonological units back to word activations. Effects of this kind were demonstrated in the TRACE architecture (which in fact had its phonological layer between words and phonetics in the traditional manner) whereby phonotactics - constraints on legal phoneme sequences - became expressed in the phonological layer as a consequence of the links to word activations.

Phonological Sequence Units. So far the knowledge that we have exploited has concerned discrimination between lexical entries at an instant of time. The other important knowledge sources are those that provide constraints on the development of activations of lexical entries and phonological units over time. In a simple network model which uses replications of layers to represent the time sequence of the decoded utterance, sequence constraints correspond to units which tie together word units and phonological units between layers. See Figure 2.

Phonological sequence units have two important functions: firstly they provide phonological unit ordering information on the currently activated word sequence (the combination of individual units and short sequences constrains the overall activated sequence; we do not wish to treat words as a 'bag' of unordered phonological activations), and secondly they establish a phonological context for the exploitation of knowledge about modification to phonetic realisations of words in context.

This second function of phonological sequence units, the implementation of 'phonological rules' or 'fast-speech rules', is important in the proposed architecture since there are no direct connections between phonetic representations and phonological ones. The function is described in more detail in section 4, where we investigate the interface between the phonetic component and the lexicon.

Prosodic Units. The sole use of the word layer to filter the phonetic evidence through to the phonological units effectively prevents the construction of phonological units activated by large-scale prosodic structures in the phonetic evidence. Thus in addition to the word units, we shall need units representing prosodic components: parts of an intonation contour, for example. These prosodic units will tie the output of the phonetic component over a larger timescale than words to phonological units representing a prosodic analysis of the utterance. In many respects these additional prosodic units act as word units: they compete for the interpretation of the evidence, and discrimination between different prosodic interpretation is heightened by feedback from the phonological analysis. The word units and the prosodic units are not linked directly, although the prosodic interpretation could interact with the segmental interpretation.

Grammatical Units. Sequence units representing syntactic constraints could be developed from traditional constituent-structure analysis, but equally they could be calculated from N-gram statistical analysis of corpora. Word sequences which fit acceptable patterns would have higher levels of activation, words which are required to fit a highly popular constituent will be given an activation prior to the phonetic evidence for the word being available. We might even consider how external parameters of dialogue state could pre-activate combinations of words, constituents or grammatical groups. These are details beyond the competence of the author.

In the next section we shall look at the phonetic component which activates the word units, and subsequently how the network lexicon and the phonetic component are joined. Before this, it is important to make a statement of the feasibility of the network lexicon.

The a priori construction of a network lexicon with this type of structure is not something that can be attempted for other than an extremely modest recognition task. The PDP lexicon contains an awesome number of parameters, even given the pre-definition of what the units represent. Such a lexicon will have to be constructed incrementally from continued experience with the interpretation of speech signals. ASR systems must allow growth if they are to accommodate realistic speech communication; with the network lexicon, incremental growth is the only method by which it could be constructed.

3. The Phonetic Component

There is an ironic symmetry in the problems of phonetic labelling and phonetic recognition. On the one hand, labellers of speech databases are only too aware of the compromises needed to identify abutting extents of the speech signal with a single label from a small vocabulary (Seneff and Zue, 1988). On the other hand, constructors of recognition systems appeal to 'coarticulation' to explain why their systems do not recognise these very same labels (Chow et al, 1987).

Huckvale (1990) argues that linear transcription is exploited in ASR because it is convenient for contemporary pattern recognition systems, and because it also happens to match an outdated phonological model. This section describes a multi-dimensional phonetic analysis of the signal which could be generated automatically on pattern recognition principles, since it exploits a hierarchical recognition architecture and explicitly separates phonetic from phonological representations.

Labelling. The first subject to address is the format of the phonetic representation chosen to label speech signals. Since we wish to recreate this labelling automatically, then we need a representation that can be associated with the signal in a consistent and uncompromising manner, and with no need for prior phonological knowledge.

What type of labelling would have these properties? Firstly it cannot be a single sequence of labels because phonetic parameters of the signal can change independently, e.g. obstruction and voicing. Secondly each level of representation cannot be a discrete sequence of labels because of the arbitrariness associated with localising changes in phonetic content in time (clearly phonetic elements overlap). Thirdly every level cannot be tied accurately to the time course of the signal, because evidence for a phonetic element is distributed: e.g. syllabicity.

We are drawn to a multi-dimensional phonetic feature representation which has varying degrees of temporal accuracy. Closest to the signal we can label fine temporal events: pitch periods, bursts, onsets, changes in periodicity, spectral transitions. Further from the signal, and less well-specified in time: vocalic portions, obstruent transitions, frication types. And at the highest levels some prosodic interpretations: stress patterning, syllabic nucleii, pitch accents, with still coarser temporal specification. See Figure 3.

Outputs. There are also important properties a phonetic component output must have in ASR:

a. **Speaker Normalisation:** The phonetic component output should not contain vocal-tract specific information (such as absolute formant frequencies). Thus it must make use of a parametric description of the speaker in the transformation of the signal.

b. **Acoustic Normalisation:** The phonetic output should make phonetic judgements about the signal in a variety of acoustic conditions (the better the acoustic environment the more specific and more reliable the phonetic outputs). Thus it must make use of parametric assessments of the acoustic environment and channel.

c. **Variability as Probability:** The phonetic output should represent the probability of each phonetic event at a given time from an input containing a variety of acoustic realisations of the event. Thus acoustic variability must be modelled and exploited much as a Gaussian classifier measures the probability that a given pattern vector comes from the same population as a set of training vectors.

Recognition. The third aspect of the phonetic component is the architecture for pattern recognition that might be trained to recreate this kind of labelling. The first point here is that it cannot be specified in advance whether the multi-dimensional feature traces are derivable from each other (in a hierarchy) or whether they need direct access to the signal as well (in a heterarchy). Can 'syllabicity' be derived solely from more simply feature descriptions, or is it a different type of property of the signal? Thus the safest choice for recreating the labelling hierarchy is to construct a recognition heterarchy. In terms of a feed-forward network (such as the multi-layer perceptron) we would construct a pattern classifier that took as input a window on the speech signal and output the feature labels. However the internal structure of the network could relate to our belief in the hierarchical structure of the labels, with the outputs being formed at different levels in the network and with each level in the network having access to every lower layer as well as the input directly, see Figure 4. Each network layer would also require hidden units that maintained internal representations.

The combination of labelling hierarchy and recognition heterarchy has greater potential than either alone: the labelling hierarchy needs more than a sequence of independent transformations to process it, the recognition heterarchy could not be trained without hierarchically labelled material.

What is the feasibility of constructing such a phonetic component? Once again, to construct a complete system from scratch would be too large a task, and we must consider 'bootstrapping' methods. Starting with many repetitions of simple hand-labelled utterances, and growing to more complex utterances using semi-automatic methods of labelling (some early work in this regard is described in Huckvale et al, 1989). One other important point is that the performance of the phonetic component cannot be assessed in isolation from some task to which it might be put: in phonetic recognition or in time-alignment of transcription, for example. A performance figure of 95% for some feature is meaningless when separated from the consequences of that performance for some task. Howard and Huckvale (1989) describe the use of a feature-based front-end to an isolated word recognition system, where task performance is shown to be sensitive to the performance one particular feature detector.

4. Joining Phonetic Component to the Lexicon

Section 2 has described a theoretical construct: a network lexicon that incorporates phonological and syntactic analysis. Section 3 has described a phonetic component constructed using known pattern recognition methods which outputs phonetic feature probabilities from an analysis of the signal. Clearly we need to marry these two, indeed it is this junction that embodies the key signal-to-symbol transformation.

There are a number of aspects to the junction:

a. **Word-unit activation:** The word units in the lowest lexicon layer need to be activated with the probability that such a word is present in the signal at a given position in time. Thus we need a model of the phonetic realisation of the word as a function of time, which suggests existing recognition techniques for whole-word pattern matching applied to the output of the phonetic component. The output of the whole-word model could be the best match between the model and the signal for all starting times earlier than the current time (as in word-sequence recognition, Bridle et al, 1982).

b. **Word-model training:** Initial set-up of the word models could be made from a corpus of words or from predictions of existing phonetic recognition systems. There must, however be a mechanism for continued development of the models with experience in recognition and as a consequence of recognition errors. Clearly there is a link between development of phonological representations in the lexicon and the increased sophistication of the word models.

c. **Alternative word realisations:** some variability in word unit realisation is accommodated by the word model. However, just as now with acoustic models, some variety falls outside convenient statistical parameters, and additional models have to be constructed. Thus idiosyncratic pronunciations of words ('bath' as /ba:th/,/bath/), major simplifications in connected speech ('and' as /n/), or phonetic choice (e.g. stress shifting) can be incorporated as additional, separate phonetic models linking to replications of the word unit. The need for additional models can be determined by established techniques such as clustering, or better through experience with recognition, so that only the fewest pronunciation alternatives are used to meet lexical discrimination requirements.

d. **Contextual dependency:** Word model varieties are of course related to the phonological context, and the system needs to have a mechanism for representing regularities between phonetic variation and phonological unit activation. These regularities are sometimes known as 'phonological rules' or 'fast-speech rules' (Oshika et al, 1975). Thus the (phonetic) elision of alveolar stops can be seen to be related to a phonological context involving a preceding fricative and a following consonant: e.g. 'next week' as /neks wi:k/. The recognition system needs not only to allow for this variety in the word 'next' but to establish the regularity by which all word sequences containing [fricative] [alveolar stop, same voicing as fricative] [consonant] can cause the elision of the stop. As introduced in section 2, phonological sequence units can be activated by selected phonological contexts, and hence the potential context for a fast-speech rule can be detected and represented. The activation of this context can then be used to support alternative word realisations that differ according to this rule. Consider Figure 5: the phonetic evidence of /neks wi:k/ activates words 'necks', 'nex[t]' and 'week' equally and 'next' less so. Phonological units representing /k/, /s/ and /w/ are activated from these word activations which in turn activate a phonological sequence unit representing the context of the alveolar stop elision rule. This gives additional weight to the /t/ hypothesis (activation), which in turn supports the 'next' word unit.

This description of the relation between the phonetic component and the network lexicon, while still needing much more development, does maintain a consistent theoretical position, which is the separation of phonological representations from phonetic ones. We have avoided the temptation to connect the output of the phonetic component directly to phonological units. Only experimental work with such a model will show whether such a strong theoretical position is a help or a hindrance in speech recognition.

5. Conclusions

I have sketched out a Parallel Distributed Processing (PDP) architecture for speech recognition that tries not to compromise between linguistic theory and practicality. Experimental work at University College has only addressed one small part of this architecture. As a consequence many practical details still need to be developed: how to deal with time sequences, how to develop the lexicon incrementally, how best to label signals and train pattern recognition schemes for the phonetic component, how to control construction of the phonetic/lexicon junction, and how to find optimal solutions in a mixed Neural Network/statistical pattern matching system. Similarly there is the need for theoretical analysis of the best choice of phonological representations.

Advances in speech recognition will only come through the exploitation of our understanding of speech communication as expressed in our formalised knowledge. Released from the linear phonological model, exploiting a multi-dimensional phonetic representation, and maintaining the separation of phonology and phonetics, the architecture described in this paper demonstrates the innovative power of PDP.

One final word of warning, I have set out to exploit existing speech knowledge in ASR, not develop a cognitive model of speech recognition: whilst there should be similarities between the word lattice produced by this system and the word sequence produced by a human listener, the system should not be judged on whether it reproduces other characteristics of human speech perception.

Acknowledgement

The work described in this paper is supported by a Fellowship from the Science and Engineering Research Council of Great Britain.

References

J.S. Bridle, M.D. Brown, R.M. Chamberlain (1982), "A one-pass algorithm for connected word recognition", Proc. IEEE ICASSP-82, Paris, pp899-902.

Y.L. Chow, M.O. Dunham, O.A. Kimball, M.A. Krasner, G.F. Kubala, J. Makhoul, P.J. Price, S. Roucos, R.M. Schwartz (1987), "BYBLOS: the BBN continuous speech recognition system", Proceedings ICASSP-87, Dallas, p89-92.

I.S. Howard and M.A. Huckvale (1989), "Two-level recognition of isolated words using neural networks", Proceedings IEE conference on Artificial Neural Nets, London, pp90-94.

M.A. Huckvale. I.S. Howard, W.J. Barry (1989), "Automatic phonetic feature analysis of continuous speech", proc. EuroSpeech-89, Vol2, pp565-568.

M.A. Huckvale (1990), "Exploiting Speech Knowledge in Neural Nets for Recognition", Speech Communication 1990 part 2.

T. Kohonen, K. Torkkola, M. Shozakai, J. Kangas (1987), "Microprocessor implementation of a large vocabulary speech recognizer and phonetic typewriter for Finnish and Japanese", Euro. Conf. Speech Tech. Edinburgh, Vol2, pp377-80.

K.F. Lee, H.W. Hon, M.Y. Hwang, S. Mahajan, R. Reddy (1989), "The SPHINX speech recognition system", Proceedings ICASSP-89, Glasgow, p445-448.

S.E. Levinson (1985), "A unified theory of composite pattern analysis for automatic speech recognition", in Computer Speech Processing, ed F. Fallside and W.A.Woods, Prentice Hall.

J.L. McClelland, and J.L. Elman (1986), "Interactive processes in speech perception: The TRACE model", in Parallel Distributed Processing: Explorations in the Microstructure of Cognition, ed. by D.E. Rumelhart, and J.L. McClelland, MIT Press, Vol 2, Chapter 15.

B. T. Oshika, V.W. Zue, R.V. Weeks, H. Neu, J. Aurbach (1975), "The role of phonological rules in speech understanding research", IEEE Transactions ASSP 23 p104.

S. Seneff and V.W.Zue (1988), "Transcription and alignment of the TIMIT database", NIST TIMIT database documentation.

A. Waibel, H. Sawai, K. Shikano (1989), "Consonant Recognition by modular construction of large phonemic time-delay neural networks", Proceedings ICASSP-89, Glasgow, p112-115.

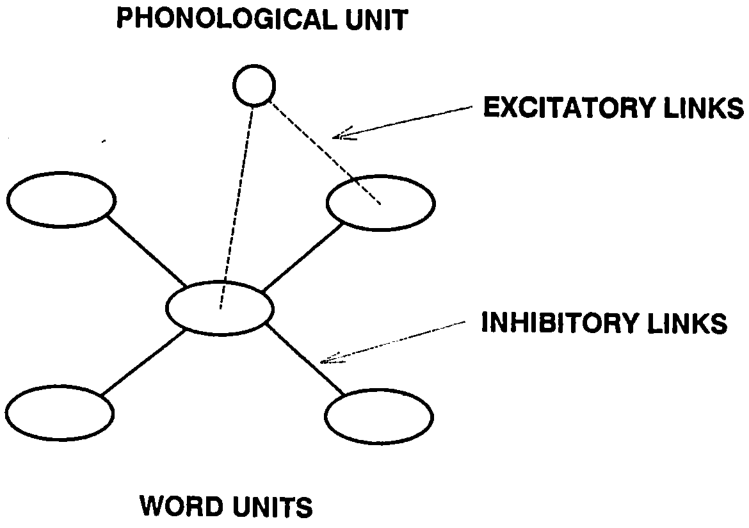

Figure 1. Phonological Units: These tie together word units that share the same phonological prescription. The word units are activated from below by a phonetic component and compete to explain portions of the signal.

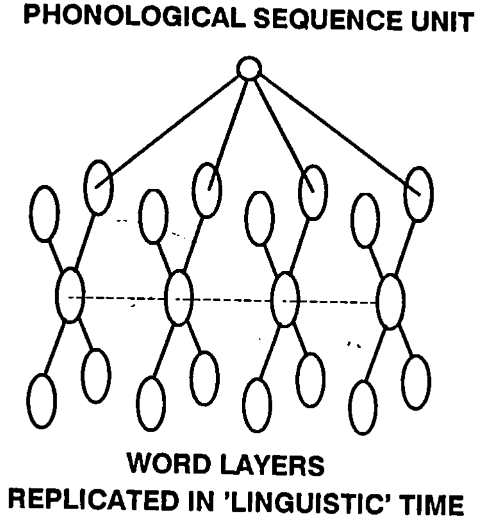

Figure 2. **Phonological Sequence Units:** These tie together sequences of word or phonological units by linking units across layers that represent development of the linguistic content of the message.

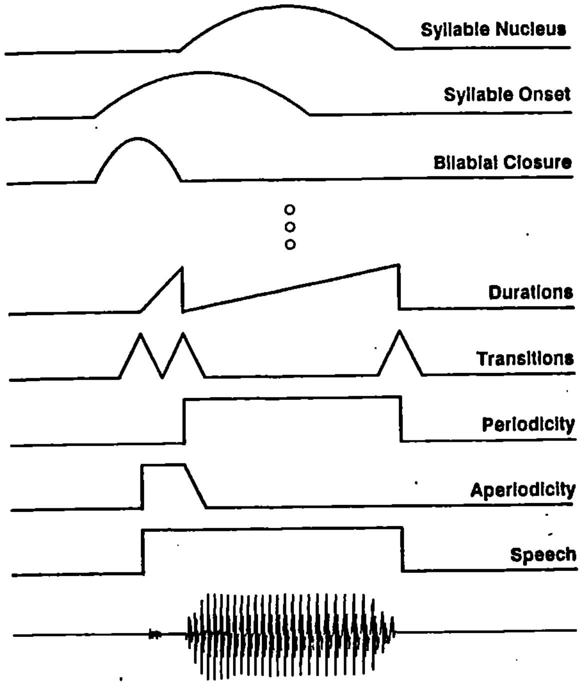

Figure 3. **Multi-dimensional Phonetic Labelling:** A speech signal (here /pa/) can be labelled simultaneously at a number of levels. Each level describes some acoustic or phonetic aspect of the signal. The hierarchy allows events to be placed in context, with appropriate degrees of temporal accuracy avoiding the compromises of linear labelling.

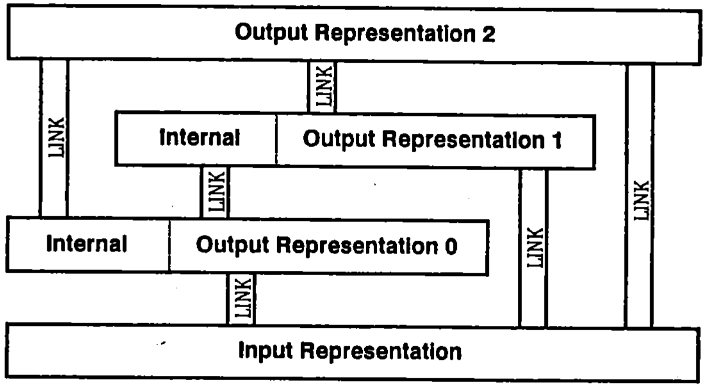

Figure 4. **Heterarchical Classifier Structure:** Each layer in a feed-forward network can have access to all earlier layers. There are outputs of the system at all layers, which are used for training with hierarchically-labelled material. There are also hidden units at each layer so that the network can develop internal representations to tie layers together.

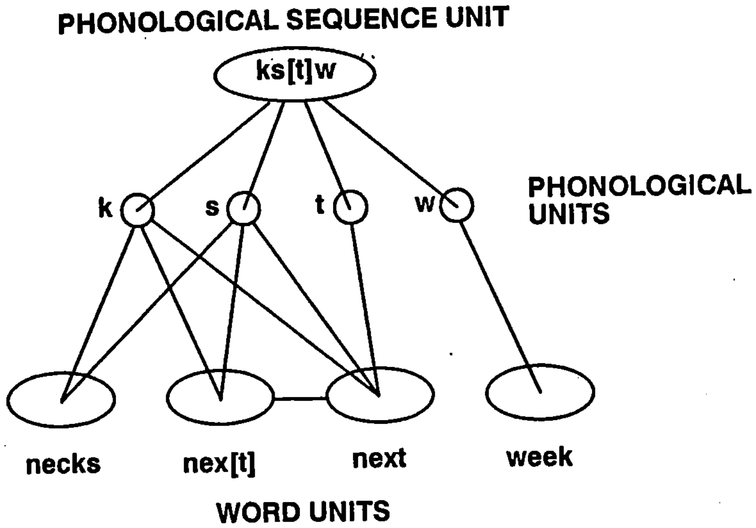

Figure 5. **Phonological Rules:** Phonological sequence units can implement 'fast-speech' rules by detecting the context in which effects occur and then feeding back to word alternatives. Here the phonetic sequence /neks wi:k/ activates "necks" and "nex[t]" equally, but the phonological sequence unit representing /ks[t]w/ (the elision of /t/ in context) supports the activation of "next".

IMPLEMENTATIONS OF AN INTONATION ALGORITHM FOR SYNTHESIS-BY-RULE

Mike JOHNSON

SPEECH, HEARING AND LANGUAGE: WORK IN PROGRESS
U.C.L. No 4 (1990)