EXPLOITING SPEECH KNOWLEDGE IN NEURAL NETS FOR RECOGNITION

Mark HUCKVALE
Department of Phonetics and Linguistics, University College London, Gower Street, London WCIE 6BT, U.K.

Received 1 December 1989

Abstract. This paper argues that neural networks are good vehicles for automatic speech recognition not simply because they provide non-linear pattern recognition but because their architecture allows the incorporation and exploitation of existing knowledge about speech. The paper is in two parts: Part I defends the need for the incorporation of existing knowledge, while Part II sketches a speech recognition architecture that uses neural networks to represent and exploit existing phonological and linguistic knowledge. The first part of the paper argues that the definition of the speech recognition problem implies that prior knowledge of linguistic analysis is essential for its solution, and suggests that the currently poor exploitation of such knowledge is a consequence of contemporary pattern recognition architectures. Criticism is made of the current emphasis on syntactic pattern recognition algorithms operating at the level of the phonetic segment. The second part of the paper demonstrates that a network architecture for the lexicon provides a mechanism for the incorporation and exploitation of a range of phonological analyses. Furthermore, through the explicit separation of phonological representations from phonetic ones, there exists the possibility of constructing a front-end phonetic component on purely pattern recognition principles. Through normalisation of speaker and environment, the phonetic component may be interfaced to the network lexicon to provide a complete recognition architecture which avoids compromise in the exploitation of speech knowledge.

PART 1

The need to exploit existing knowledge in automatic speech recognition systems

Much of the recent published work that combines neural network processing architectures and speech recognition has been concerned with how the non-linear pattern recognition capabilities of networks provide alternative or better recognition performance on standard speech recognition tasks (e.g. Waibel et al., 1989; Kangas and Kohonen, 1989). This paper argues that just as there is more to speech recognition than recognising patterns, there is more to neural networks than pattern recognition. The central proposition of this paper is that neural networks are a natural formalism for the expression and exploitation of linguistic and phonological knowledge, and it is only through the use of existing knowledge about speech that the fundamental problems of speech recognition can be addressed.

This paper is in two parts: Part I deals with how "speech knowledge" (a general term for acoustic-phonetic, phonological and linguistic knowledge) is currently exploited in speech recognition systems, concentrating on the use of the linear phonological model. The aim of Part I is to show that speech knowledge must be exploited to overcome the limitations of existing systems, whether you believe (as I do) that such knowledge is central to the formulation of the problem, or whether you believe that to learn speech from scratch would simply take too long. While the discussion in Part I is pertinent to both conventional and neural network recognition systems, Part II of the paper is specifically concerned with neural network architectures for the exploitation of phonological and acoustic-phonetic knowledge. Part II describes a network lexicon which can embody different models of phonological analysis (a construct developed from the TRACE model of speech perception (McClelland and Elman, 1986)) and a phonetic component built on a feedforward perceptron classifier which avoids compromise in the labelling of speech signals.

1. Speech knowledge in contemporary systems

What does a computational device need to know about speech in order to recognise it? Let us start with a general definition of Automatic Speech Recognition (ASR) as:

> *the automatic generation of a shallow-enough word-lattice for a given task from a spoken utterance and some information about the speaker, environmental conditions, communication channel and dialogue state.*

Where "shallow-enough" depends on how much domain-specific knowledge is available for resolving the word sequence and the requirements for acceptable performance. Thus the knowledge a system needs for recognition, on this definition, is the relationship between signals and words and *how external parameters* of speaker, environment, channel and dialogue state affect that relationship. See Fig. 1.

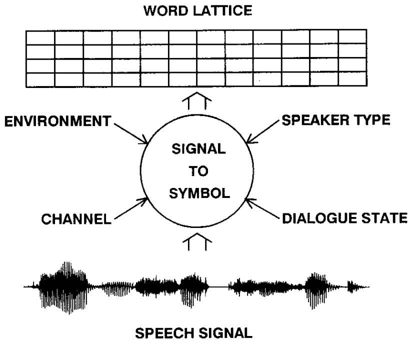

Fig. 1. *Speech recognition task:* The mapping from signals to symbols is conditioned by the setting of external parameters of acoustic environment and channel, speaker type and dialogue state.

Real systems are specialisations of this definition. Some specialisations are chosen to make ASR easier: only one speaker, only one environment/channel, a small task vocabulary, restrictions to isolated words. Lack of knowledge about the acoustic formulation of words or the influence of external parameters is replaced by control over the operating characteristics of the device.

Other specialisations appear to make ASR more difficult: multiple speaker recognition with no information of speaker type, a range of environments and channels with no information of acoustic conditions, systems with no dialogue model, treatment of continuous speech as concatenated isolated words. Lack of information about external parameters or weaknesses in the acoustic modelling of word sequences is often re-placed by flexibility of pattern matching: a model of ignorance" (Makhoul and Schwartz, 1985) that accommodates unpredictable variation.

Designers of this second set of systems do not choose such specialisations out of perversity, but because of the difficulty of constructing ASR architectures which allow for the exploitation of context/speaker/environment/dialogue knowledge in the recognition process. Thus, although some parameters of a speaker could be extracted from a signal (e.g. fundamental frequency distribution or estimate of vocal tract length) conventional speaker-independent systems (e.g. DeMichelis et al., 1989) do not use pattern recognition techniques which could exploit this information. Similarly, acoustic processing in telephone based recognition systems aim to normalise the signal (e.g. Imamura et al., 1989), rather than provide the recognition process with information about the channel characteristics.

The important point is not that we lack knowledge about how external parameters might influence the decoding of signals or how context influences phonetic modifications to words, but rather we lack architectures which could exploit that knowledge. But before we look in more detail at this latter issue in section 2: the exploitation of acoustic-phonetic and phonological knowledge, let us first look at what speech knowledge is exploited in contemporary ASR systems and comment on some popular misconceptions.

(a) Existing systems do not use speech knowledge anyway.
Impressive technical achievements such as the Tangora dictation system (Bahl et al., 1989) or the Sphinx speaker-independent system (Lee et al., 1989) may not be constructed around state-of-the-art linguistic theory, nor use the latest phonological analysis of English, but nevertheless they are built on several decades of slow but steady experimental analysis of speech and language. The word models of the Tangora system are built from sequences of phonemic models, Sphinx exploits the differences in variability between function words and content words , both use N-gram models of syntax. All recognition systems assume the existence of words, that signals can be analysed as a sequence of words, that word sequences can be determined without knowledge of what the speaker intended by producing the utterance. These hypotheses about speech derive from prevalent theoretical models of speech communication. The existence of phonemes, words, sentences and intent is part of our analysis of language that these systems exploit. The real issue is why these systems use such little knowledge from such weak and old-fashioned theoretical models, and one answer is that contemporary computational architectures place restrictions on what can be effectively exploited (see section 2). It is more important to use a little knowledge well than a lot of knowledge badly.

(b) Speech is just a pattern recognition problem.
The "engineering view" of speech recognition holds that the relationship between signals and external parameters on the one hand, and symbols on the other, can be learned from a set of labelled pattern vectors. Thus there is no need for the incorporation of *a priori* speech knowledge, since all such linguistic and phonological analysis is itself only derived from the signal and can therefore be induced by the system. If a child can learn language by example, then so can a machine.

It is important to make a direct attack on this proposition which confuses what is possible in principle with what is possible in practice. Firstly, *a priori* knowledge of relationships between speaker characteristics, environment/channel, context and dialogue state and the interpretation of the signal as symbols is essential because it reduces to a manageable size the volume of training material necessary to explore the influence of external parameters. To analyse, say, 100 sub-word units from 100 speaker-types, in 100 acoustic environments, in 100 linguistic contexts for 100 repetitions without making assumptions of the independence of speaker-type from environment, etc., the blind pattern recognition system would require 10^{10} pattern vectors! However, to exploit the independence of external parameters implies an architecture with *a priori* internal representations. Thus in the IBM dictation system, where there was too little textual material to accurately analyse the statistical properties of trigram word models, grammatical analysis of words was used to aid the probability estimation, which in turn brought in *a priori* knowledge of word classification (Jelinek, 1985).

The second attack on the "speech is pattern recognition" proposition is more subtle, but directly aimed at systems which aim to produce, by pattern classification, a sequence of sub-word linguistic units (phonemes or demi-syllables, etc.). Clearly, speech is not *produced* as a sequence of phonemes nor a sequence of any other phonological representation; speech is simply the sounds a talker makes when he/she uses language with the intent of changing the mental state of a listener. It is the analysis of these sounds in the context of the meaning of the message which gives rise to units such as words, syllables and phonemes. While these units must be related to how speech and language is encoded and processed in the brain, they are only indirectly related to how speech is encoded and processed as a signal. If phonological units are cognitive at all they only emerge as a consequence of processing the signal in the brain. A moving image on a video screen is not to be found directly in the video signal: it is only through processing the signal into lines, then frames, then exploring changes between frames that the movement emerges. The communication of the moving image/linguistic unit relies on the structure of the encoder and interaction of the signal with the structure and knowledge of the decoder. Speech recognition without knowledge of linguistic analysis is like decoding television broadcasts with a crystal radio set.

(c) Speech knowledge implies hard constraints.

Of the rather few advances in speech recognition science made over the past thirty years, "delayed decision making" has proved a powerful and practical principle. The idea has roots in the use of Dynamic Programming as a search strategy in Harpy (Lowerre and Reddy, 1980), and John Bridle expounded it clearly in a connected word recognition algorithm in which the words and the word sequence were determined simultaneously (Bridle et al., 1982). The principle is currently to be seen in the HMM-based large-vocabulary speech recognition systems, where decisions of the recognised phone sequence can only be made after the determination of the best word sequence. In a different paper (Huckvale, 1987) 1 have argued that all speech recognition systems can be viewed as having an implicit model of speech production and an algorithm that determines an input to the model given a hypothetical output (an unknown utterance). Since we would like to choose the most likely input to our production model given the unknown signal, we turn to known optimal search methods such as Dynamic Programming.

There is one important consequence of this view of recognition as optimisation: namely that we do not use constraints on speech production to "filter" representations as they are processed up the decoding hierarchy, rather constraints are used simultaneously to derive an explanation. For example, we could consider a phonotactic constraint that syllables do not begin with /gm-/, but to exploit this bottom-up would be to reject legal phoneme sequences that had erroneous syllabification. Using such a rule as a constraint in an optimisation-based system would not cause a problem since syllabification itself is constrained to fit the production model. The result is an interpretation that fits the constraints without using any single constraint earlier than another.

To summarise, existing systems do use speech knowledge, but only as much as can be exploited effectively; the nature of the speech decoding problem is that we need to use human pre-conceptions about the structure of the signal; and lastly that the exploitation of knowledge can be incorporated safely if performed in the right computational framework. Part II of this paper takes these three ideas one step further in the discussion of neural network processing architectures for speech recognition. We look at how network architectures can incorporate modern phonological models, how experience with language can be used to construct a phonetic recognition scheme, and how networks as "constraint satisfaction" systems can perform recognition on the basis of the interaction of phonetic evidence and stored speech knowledge, if not as yet in an optimal way.

2. A criticism of linear phonology in ASR

Of all the types of existing knowledge that are currently exploited in ASR, that resulting from phonological analysis of speech is most widespread. This section explores contemporary use of phonological knowledge in ASR and makes some criticisms. This section acts as a bridge to an alternative network representation of phonology in Part II.

All existing systems with vocabularies of > 1000 words interpose a phonological representation between the signal and the word lattice (e.g. Tangora (Bahl et al., 1989), Byblos (Chow et al., 1987), Sphinx (Lee et al., 1989), and Kohonen et al., 1987)). This is because (i) users cannot be expected to speak thousands of words simply to train such systems, and (ii) independent word models do not provide sufficient discrimination between similar words. A phonological layer is required to structure the acoustic differences and similarities between words.

Thus, instead of making separate models of the words "pin" and "bin", models are made of "p-" (or "pi-") and "b-" ("bi-") as well as "i" and "n" The complete acoustic models for the words then have identical second halves, forcing discrimination onto the first half of the words - the parts of the words that phonological analysis of the language has shown to demonstrate the difference between them.

In fact, existing systems make use of this interposed phonological representation in two markedly different ways. In the phonetic transcription systems of Waibel et al. (1989) and Kohonen et al. (1987), the phonological units are used as a representation in the recognition process: the signal is transformed to a sequence or lattice of phonological symbols. In the statistical models of Tangora, Byblos and Sphinx, the phonological units are only used to structure the acoustic similarities between words during training; during recognition the determination of the phonological symbol sequence/lattice is subsumed into the recognition of the word sequence/lattice. In these systems, the phonological units are simply a means for incorporating knowledge about speech into the recognition knowledge base.

The need for phonological knowledge of this kind is widely accepted as a means for addressing the ASR task specified in section 1. However, it is important to realise that the choice of phonological transcription - a linear symbol sequence - was not made because of the superiority of its corresponding (linear) phonological theory. A choice of linear transcription is made in these systems because of its correspondence with the architecture of contemporary pattern recognition systems. The recognition architecture of Kohonen et al. (1987) converts signals first to a path through a two-dimensional map, then to a sequence of phonological symbols, then to a sequence of letters. As applied to ASR, Hidden Markov Models are implementations of finite state parsers: they seek an interpretation of the signal in terms of a single path through a network of states.

Where these systems do exploit knowledge of the hierarchical nature of speech representations, it always reduces to constraints on phonetic symbol sequences. Thus Kohonen's system uses constraints on the phonological symbol to letter sequence mapping, Tangora exploits lexical and syntactic constraints to construct a network of phone models.

Even the traditional phonological analyses of speech (e.g. Gimson, 1989) use a multi-dimensional phonetic representation of two sequences: the segmental and the supra-segmental, reflecting exploitation of the large degree of independence in the speech production mechanism between articulation and air-flow/voice-source. More recent analyses (see e.g. Lass, 1984, ch. 10) use even more "layers" of phonetic representation, primarily organised around the syllable, which can provide more parsimonious descriptions of phonetic constraints and realisations.

Before we consider the potential for the exploitation of non-linear phonological analysis in section 3, and the utility of a multi-dimensional phonetic representation in section 4, let us look at the weaknesses of the linear phonetic model in explaining the acoustic form of the speech signal:

(a) *Segmentation:* linear units are not readily segmented from the sound stream. E.g. the /b/ in "bin" is not a separate entity from the vowel.

(b) *Locality of acoustic evidence:* linear units make assumptions about the locality of evidence for discrimination of units. E.g. the acoustic evidence for /b/ in "cabin" does not reside solely in the stop release and transition.

(c) *Contextual dependency:* the acoustic realisation of linear units is not independent of the selection of adjacent units. E.g. the acoustic form of /b/ varies before different phonological vowels.

(d) *Weak sequence constraints:* the linear model assumes the freedom to produce arbitrary unit sequences. E.g. syllable initial /bn-/ is as possible as /bl-/.

(e) *Similarities of acoustic form:* the linear model fails to account for the structural similarities between the acoustic realisations of units. E.g. a contextual model of the realisation of /b/ does not provide any information about the contextual realisation of /p/.

The construction of ASR systems around the linear model can be seen to be compensating for these weaknesses: the selection of larger subword units (e.g. Yoshida et al., 1989), pragmatic augmentation of the phone inventory (Lee et al., 1989), context-dependent phone realisation (Chow et al., 1987), and the general delay in phonetic decision taking (see section 1). The consequences of the linear model for ASR are severe: (i) difficulty in labelling training material and reconstructing the labelling by pattern recognition (e.g. Cole and Hou, 1988), (ii) inadequate phonetic discrimination of simple words (e.g. E-set, Bedworth et al., 1989), (iii) complexity of using connected speech (Bahl et al., 1989), (iv) need for large quantities of training material (Chow et al., 1987).

Without suggesting that non-linear phonetic and phonological models alone will solve these problems, it is clear that the continued use of the linear model is holding up the development of speech recognition systems. But because of the correspondence with pattern recognition architectures, the stranglehold of the linear model will be hard to break.

The answer is not to abandon phonology though, and to explore acoustic segmentation (as in Zue et al., 1989), or to build systems that derive their own phonological analysis (Svendsen et al., 1989), but to use the knowledge we have in an appropriate way. This means considering carefully the role each piece of speech knowledge that we have can play in a speech recognition system, and devising computational architectures that can exploit it. This leads us, conveniently, to a discussion of neural networks.

PART II

Speech knowledge & network architectures

3. A network lexicon for ASR

Although we are still in the infancy of a mathematical analysis of the behaviour and optimisation of parallel distributed processing (PDP) architectures, the PDP philosophy of distributing knowledge through a network of interacting units has a synergistic relation with the problems of linear phonology in ASR.

Specifically, let us consider the lexicon as a network rather than as an unordered set of word specifications, each containing separate pronunciations and grammatical attributes. The description below is a direct descendant of the lexical layer in the TRACE model of speech perception (McClelland and Elman, 1986).

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 4 and 5) 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 *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 - using one layer, or in space - 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. As mentioned in section 2, 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, and many others. These units have positive bidirectional connections with the word units. Thus if "bin" is activated, then some of that activation is shared with 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 Fig. 2.

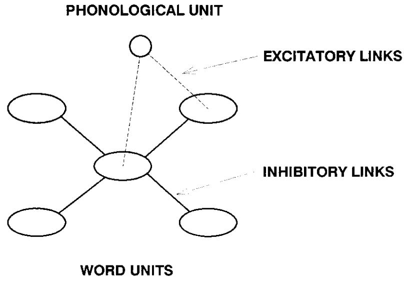

Fig. 2. *Phonological units:* These tie together word units that share the same phonological prescription. The word units are activated from below by a phonetic component and compete to explain portions of the signal.

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.

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 Fig. 3.

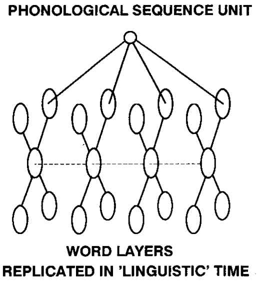

Fig. 3. *Phonological sequence units:* These tie together sequences of word or phonological units by linking units across layers that represent development of the linguistic content of the message.

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 5, where we investigate the interaction 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 can 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 scope of this paper and have been dealt with by other authors (e.g. Hinton, 1981).

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.

4. A phonetic component for ASR

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).

Section 2 argued that linear transcription was exploited in ASR because it was convenient for contemporary pattern recognition systems, and because it also happened 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, 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 Fig. 4.

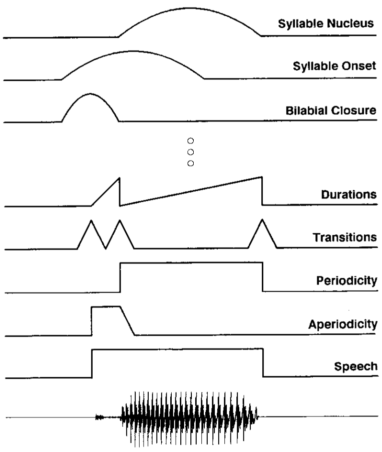

Fig. 4. *Multi-dimensional phonetic labelling:* A speech signal (here /pa/) can be labelled simultaneously at a number of levels. Each level describes some acoustic or phonetic aspect of the signal. The hierarchy allows events to be placed in context, with appropriate degrees of temporal accuracy, avoiding the compromises of linear labelling.

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 each 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 Fig. 5. Each network layer would also require hidden units that maintained internal representations.

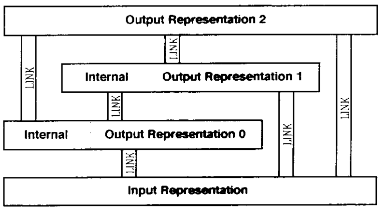

Fig. 5. *Heterarchical classifier structure:* Each layer in a feed-forward network can have access to all earlier layers. There are outputs of the system at all layers, which are used for training with hierarchically-labelled material. There are also hidden units at each layer so that the network can develop internal representations to tie layers together.

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 of one particular feature detector.

5. Joining phonetic component to the lexicon

Section 3 has described a theoretical construct: a network lexicon that incorporates phonological and syntactic analysis. Section 4 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 /bɑːθ/, /bæθ/), 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 4, 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 Fig. 6: 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.

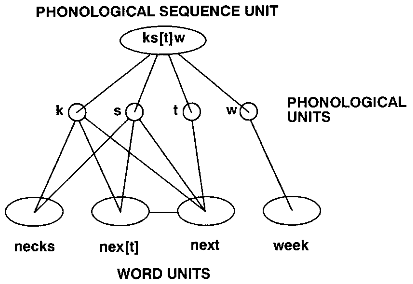

Fig. 6. *Phonological rules:* Phonological sequence units can implement "fast-speech" rules by detecting the context in which effects occur and then feeding back to word alternatives. Here the phonetic sequence /neks wi:k/ activates "necks" and "nex[t]" equally, but the phonological sequence unit representing /ks[t]w/ (the elision of /t/ in context) supports the activation of "next".

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.

6. Conclusion

In Part II we have sketched out a 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. It is not, however, *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.

Too much research with neural networks only makes use of their pattern recognition capabilities. However, it is the way their architecture allows the exploitation of existing knowledge that makes them most promising for speech recognition.

Acknowledgement

This work is supported by a fellowship from the U.K. Science and Engineering Research Council.

References

L.R. Bahl, P.F. Brown, P.V. de Souza, P.S. Gopalakrishnan, F. Jelinek and R.L. Mercer (1989), "Large vocabulary natural language continuous speech recognition", Proc. *ICASSP-89, Glasgow.*

M.D. Bedworth, J.S. Bridle, L. Flynn, K.M. Ponting, L.Y. Bottou and F. Fogelman (1989), "Comparison of neural and conventional classifiers on a speech recognition problem", *Proc. IEE Conf. Artificial Neural Networks, London,* pp. 86-89.

J.S. Bridle, M.D. Brown and R.M. Chamberlain (1982), "A one-pass algorithm for connected word recognition", *Proc. IEEE ICASSP-82, Paris,* pp. 899-902.

Y.L. Chow, M.0. Dunham, O.A. Kimball, M.A. Krasner, G.F. Kubala, J. Makhoul, P.J. Price, S. Roucos and R.M. Schwartz (1987), "BYBLOS: The BBN continuous speech recognition system", *Proc. ICASSP-87, Dallas*, pp. 89-92.

R.A. Cole and L. Hou (1988), "Segmentation and broad classification of continuous speech", *Proc. IEEE ICASSP88,* pp. 453-456.

P. DeMichelis, L. Fissore, P. Laface, G. Micca and E. Piccolo (1989), "On the use of neural networks for speaker independent isolated word recognition", *Proc. ICASSP-89, Glasgow,* pp. 314-317.

A.C. Gimson (1989), *An introduction to the pronunciation of English,* 4th edition (Edward Arnold, London).

G.E. Hinton (1981), "Implementing semantic networks in parallel hardware", in *Parallel Models of Associative Memory,* ed. by G.E. Hinton and J.A. Anderson (Erlbaum, Hillsdale, NJ).

I.S. Howard and M.A. Huckvale (1989), "Two-level recognition of isolated words using neural networks", *Proc. IEE Conf. Artificial Neural Nets, London,* pp. 90-94. M.A. Huckvale (1987), "ASR beyond HMM", *European Conf. Speech Technology, Edinburgh,* pp. 231-234.

M.A. Huckvale, I.S. Howard and W.J. Barry (1989), "Automatic phonetic feature analysis of continuous speech", *Proc. EuroSpeech-89,* Vol. 2, pp. 565-568.

A.Imamura, H. Hamada and R. Nakatsu (1989), "Speaker-independent word recognition through telephone networks using hidden Markov models", *Proc. EuroSpeech*89, Vol. 1, pp. 171-174.

F. Jelinek (1985), "The development of an experimental discrete dictation recogniser", *Proc. IEEE,* Vol. 73, pp. 161-164.

J. Kangas and T. Kohonen (1989), "Transient map method in stop consonant discrimination", *Proc. EuroSpeech-89, Paris,* pp. 345-348.

T. Kohonen, K. Torkkola, M. Shozakai and J. Kangas (1987), "Microprocessor implementation of a large vocabulary speech recognizer and phonetic typewriter for Finnish and Japanese", *European Conf. Speech Technology, Edinburgh,* Vol. 2, pp. 377-380.

R. Lass (1984), *Phonology* (Cambridge University Press, Cambridge).

K.F. Lee, H.W. Hon, M.Y. Hwang, S. Mahajan and R. Reddy (1989), "The SPHINX speech recognition system", *Proc. ICASSP-89, Glasgow,* pp. 445-448.

B. Lowerre and R. Reddy (1980), "The Harpy speech understanding system", in *Trends in Speech Recognition, ed.* by W. Lea (Prentice Hall, Englewood Cliffs, NJ).

J. Makhoul and R. Schwartz (1985), "Ignorance modelling", *in Variability and Invariance in Speech processes,* ed. by D. Perkell and D. Klatt (Erlbaum, Hillsdale, NJ).

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, Cambridge, MA), Vol. 2, Ch. 15.

B.T. Oshika, V.W. Zue, R.V. Weeks, H. Neu and J. Aurbach (1975), "The role of phonological rules in speech understanding research", *IEEE Trans. ASSP,* Vol. 23, p. 104.

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

T. Svendsen, K.K. Paliwal, E. Harborg and P.O. Husoy (1989), "An improved sub-word based speech recogniser", *Proc. ICASSP-89, Glasgow,* pp. 108-111.

A. Waibel, H. Sawai and K. Shikano (1989), "Consonant recognition by modular construction of large phonemic time-delay neural networks", *Proc. ICASSP-89, Glasgow,* pp. 112-115.

K. Yoshida, T. Watanabe and S. Koga (1984), "Large vocabulary word recognition based on demi-syllable hidden Markov models using small amounts of training data", *Proc. ICASSP-89, Glasgow,* pp. 1-4.

V. Zue, J. Glass, M. Phillips and S. Seneff (1989), "Acoustic segmentation and phonetic classification in the SUMMIT system", *Proc. ICASSP-89, Glasgow,* pp. 389-392.