An Integrated Knowledge Base for Speech Synthesis and Automatic Speech Recognition

Mark Tatham

Reproduced from *Journal of Phonetics* (1985) 13, 175—188.

Speech synthesis and automatic speech recognition systems have been conceived and implemented as different types of device. Linguistic theory has suggested that an essential property of human speech production and perception is the provision of a knowledge base, and this idea has been generally incorporated into speech synthesis systems, though only recently into automatic speech recognition systems. Developments in the cognitive theory of phonetics suggest that production and perception are mutually dependent to the extent that they can be thought of as modalities of the same system. Speech synthesis and automatic speech recognition are brought together by sharing a common knowledge base. An appropriate type of representation is outlined.

INTRODUCTION

This paper reports in outline a proposal for an integrated knowledge based system for speech synthesis and automatic speech recognition deriving in principle from ideas and metatheoretical considerations in linguistics. Speech synthesis and automatic speech recognition are seen as output and input devices for a generalised artificial intelligence system, themselves based in the principles established in that field of research.

A speech synthesis system accepts as input normal orthographic text or some computer oriented version of text (e.g. viewdata) or the language output from some artificial intelligence system. The output of the system is speech which is as near human-like as possible and which is addressed to a human being. The input to an automatic speech recognition system is human speech and the output is orthographic text or some signal to be input to an artificial intelligence system. That is, together speech synthesis and automatic speech recognition are practical simulations of the human peripheral devices for producing and perceiving speech.

In linguistics it is generally no longer argued that the human peripheral devices which produce and perceive speech are knowledge based systems, and although speech synthesis systems have always been knowledge based the same is not true of automatic speech recognition systems. It seems worthwhile discussing the place of the knowledge base in the human system as characterised in linguistics, and in the respective simulations.

THE PRESENT POSITION

Linguistics

Since the 1950s linguistics of the transformational generative school and its derivatives has modelled language behaviour in the human being as essentially a knowledge based system. Chomsky and others saw a great deal of merit in devoting the energy of linguistics research towards a symbolic characterisation of that knowledge base, rather than in continuing to develop older ideas which failed to distinguish between the knowledge base and its usage, and between symbolic and actual phenomena. The knowledge itself has been described in terms

of an abstract model enumerating certain symbolic objects and the rules constraining their co-occurrence.

Although this approach was adopted broadly throughout linguistics, as an explicit idea it penetrated comparatively late into phonetics (Tatham, 1970a). Nevertheless, it is perfectly possible to regard the phonetic component of the grammar as also being a knowledge base consisting in part of symbolic objects and rules constraining their co-occurrence, and, in addition, knowledge concerning peripheral vocal mechanisms and their manipulation. The major difference between phonology and phonetics is that phoneticians have not been concerned with cognitive aspects of speaking but with rules governing neural, mechanical, aerodynamic and acoustic systems. There is some evidence though that an important part of speech may well be cognitive (Morton, 1985), and this will be dealt with below.

Even if a cognitive aspect to phonetics is not considered it is still the case that an information base (if not a knowledge base) is required for satisfactory modelling of production and perception. The information concerns those neural, mechanical, aerodynamic and acoustic systems involved in the operation of the peripheral devices employed for speaking and hearing.

Speech synthesis systems

Contemporary speech synthesis dates from the early sixties when serious work began in what was then known as ‘synthesis by rule’ (Kelly and Gerstman, 1961; Holmes, Mattingly and Shearme, 1964), in which the focus of attention moved away from the hardware and the acoustic signal it produced to methods of control, particularly to methods of control which would permit novel utterances to be produced. It was clear that what was needed was a method of manipulating by rules a set of basic elements to produce natural sounding running speech. Ideas prevalent in phonetics in the 1960s and 1970s were incorporated, largely hinging on the fundamental notion that somehow the acoustic signal a speaker produces is the product of conjoined extrinsic allophonic segments (Ladefoged, 1971; Tatham, 1971) smoothed by mechanical and other effects.

Speech synthesis systems are therefore knowledge based. Text in orthographic form is translated into strings of phoneme symbols by reference to a set of rules specifying the relationship between orthography and phonemic representation. Phonemes are converted to allophones by rules (taken from the linguist's phonology) which describe many of the allophonic variants in normal speech. Next (though not always explicitly separated from the previous stage), allophones are conjoined by largely mechanically derived rules describing coarticulation. At this stage the rules operate on feature sets for each allophone rather than on allophone symbols — the feature sets having been obtained from lookup tables containing an entry for each allophone. The features which the conjoining rules manipulate to provide smooth transitions between segments are based on the parameters determined to be appropriate for synthesis of the acoustic waveform usually using a formant synthesizer (Witten, 1982).

Essentially speech synthesis systems are an adaptation of the general principles of the linguistics approach in phonetics and phonology (Mattingly, 1970; Tatham, 1970b). The adaptation is an insightful engineering simplification of the linguistic theory; time has demonstrated just how much of an important insight that adaptation has been. It is only after a quarter of a century that we are beginning to feel the need for more robust systems with improved naturalness which take advantage of modern cheap digital techniques and the vastly increased computational ability of the hardware.

Automatic speech recognition systems

Progress in the development of automatic speech recognition has not been as successful as that of speech synthesis. Little attention has been paid to the theory of linguistics and phonetics (with some notable, experimental exceptions (Erman and Lesser, 1980; Lowerre and Reddy, 1980)). In general the generative, knowledge based approach has not been

adopted. Most automatic speech recognition systems are little different from those of a couple of decades ago, but that the older methods are still around has not been due to the value of early insight as it was with speech synthesis, but to a dogged persistence with ideas which had been abandoned by linguistics, especially, for example, the notion that phonemes lurk somewhere in the soundwave.

Typically an automatic speech recognition system will perform some analysis of the input soundwave to extract features found to be relevant in a similar process performed passively by the human ear. The system will be designed to recognise whole, isolated words and the features for each input soundwave will be compared with stored versions of a vocabulary to find a best match. The incoming soundwave will then be declared to have been recognised (Moore, 1984). Stretches longer than single words can be recognised if treated as periods of sound analogous to single words to establish a match with stored templates themselves representing stretches of speech longer than words (Bridle, Brown and Chamberlain, 1981). Since the early 1970s the matching has been improved by time normalization between the incoming signals and the stored templates (Sakoe and Chiba, 1978), though frequency normalization (Paliwal and Ainsworth, 1985) has not produced comparable improvements in success rate. Typically such systems achieve around 95% accuracy, though this necessarily declines rapidly if the vocabulary is significantly increased.

Less often and with less success the incoming acoustic signal is segmented and an attempt made to match the segments with stored templates which may represent sub-phoneme or phoneme sized sounds or sometimes short strings of phonemes or partial phonemes. This approach has the merit of incorporating the idea that a large number of phrases can be recognised by the way in which a very small number of segments are combined. A knowledge base is required, of course, which can determine from the sequence of recognised segments which word or phrase has been input.

A major advantage of a segment oriented system is that it permits some error recovery if the rules for segment combination are sufficiently sophisticated. It is well known in phonology that the specification of features within a segment and the sequencing of segments within a word contain a certain amount of redundancy. Use of a suitable knowledge base of such redundancy rules would enable some error correction. See Cohen and Mercer (1975) for an early version of the idea that automatic speech recognition is assisted by knowledge of phonological constraints.

NEWER DEVELOPMENTS IN PHONETICS

It has recently been suggested that there may be an important cognitive aspect of phonetics distinct from the more usual cognitive considerations of phonology (Tatham, 1984). The proposed model of speech production incorporates several features which distinguish it from the speech production / perception theories which currently underpin strategies in speech synthesis and automatic speech recognition. Let us list some of the relevant notions in the cognitive theory of phonetics.

- Cognitive phonetics is formulated to interface with phonology.
- Symbolic representations are not to be confused with mechanisms (even abstract descriptions of mechanisms) or simulations. Symbolic representations are not in the system, but may be formulated from the state of the system.
- Production and perception are complementary and may largely overlap in some respects.
- At the periphery there are internally structured devices which are essentially passive in operation. In speech production gross control signals are delivered to these structures to achieve articulation.
- Fine tuning of set-piece movements is possible though there may still be observed general passive and largely mechanically dominated coarticulatory effects.
- The information enabling fine tuning comes from two sources:

- an index of semantic or phonological loading of this particular portion of the utterance,
- a scoring of perceptibility of the proposed utterance or portion of the utterance.
- Fine tuning is decided and caused actively and initiated cognitively, though it is characterised in the phonetic component of the grammar rather than in the phonology because it is to do with implementation rather than computation of phonological requirements.
- Fine tuning of articulatory structure performance is overlaid on gross control, though descriptively separate.

The cognitive theory of phonetics within linguistics attempts to characterise some aspects of speech production within the accepted metatheoretical framework. Thus a methodological distinction is made between generalizations incorporated in the speaker's knowledge base and an act of drawing on that knowledge to achieve an instance of actual speech. Because there is evidence that mechanical and other constraints on speaking are manipulated in ways which would not usually be characterised in a theory of phonology, the phonetics is claimed to have an important cognitive role drawing on knowledge of the physical constraints on the system and the limits of the controllability of such constraints. The fine tuning mechanisms exist for the systematic manipulation of these constraints for particular linguistic effects.

The idea is de-emphasized that phonetics is largely about neurophysiology, mechanics, aerodynamics or acoustics, and focus falls on mentally governed manipulation of these phenomena. Such manipulation could not take place, of course, if the speaker did not have models of their effects and did not engage in cognitive activity at a subphonological level. The theory draws on the idea of sub-systems which are essentially physical in nature (as modelled by Fowler and others (Fowler, Rubin, Remez and Turvey, 1980; Fowler, 1985)) and makes much of what might underlie the adjustment (‘tuning’ in Fowler) of the relationships internal to these sub-systems or mechanisms. To a certain extent cognitive phonetics reconciles the metatheoretical positions adopted by Hammarberg (1982) on the one hand and Fowler (1983) on the other. The position is also discussed in Parker and Walsh (1985). Cognitive phonetics is very much about the knowledge base thought to be essential for speech production and perception.

Another essential property of cognitive phonetics is that it postulates that the knowledge bases for speaking and perceiving largely overlap, to the extent *that for descriptive purposes* they might initially be thought of as the same. Indeed it is hard to imagine a production system not knowing about perception (Morton and Tatham, 1970) or a perceptual system not knowing about production. Since, if this is right, speech production must reference a model of perception and speech perception must reference a model of production the question might well in certain respects reduce to one of appropriate representation (see below).

Following notions in Fowler articulatory control is not as detailed as in more traditional theories. Two mechanisms are incorporated:

- the gross control generator and
- the fine tuning generator.

Perhaps the best way to introduce the notion of gross articulatory control which relies on internally structured low level mechanisms and overlaid fine tuning relying on some detailed cognitive activity is to proceed with some examples. There is a phonological requirement, say, for the production of a sound appropriate to match the symbolic phonological representation /a/. Among other movements of the vocal apparatus the jaw must be dropped and the tongue lowered to give the oral cavity the right shape for the necessary acoustic resonances. From some neutral position we could specify jaw dropping by *j* units of jaw movement space. And similarly we could specify the lowering of the tongue from some neutral position by *t* units of tongue movement space. This traditional full specification for these two parameters implies their independence and fails to capture their clear *inter*dependence. Dropping the jaw *j* units probably means dropping the tongue some number

of units of tongue space — though perhaps not the right number for the pronunciation of /a/. Conversely, the specification of the articulation for /i/ under similar conditions involves some jaw dropping which means dropping the tongue, but the tongue may be raised for this sound. Speakers differ as to how much they drop the jaw, but in some speakers the tongue raising must compensate for the jaw related tongue lowering tendency. Remember we are not describing observable surface phenomena ('the jaw went here, the tongue went there'), but how to *make* the jaw go here and the tongue there. So we need to know where the jaw and tongue each have to go, but need to control the movement with the knowledge of their interdependence. The theory postulates a gross control to the structure and a fine control adjusting the predicted automatic position of the tongue to the required position.

As another example we might consider the production of a soundwave appropriate to matching the symbolic representation /#ta .../ in English. We may describe a surface result of some silence, followed by some burst appropriate to the tongue's position for [t], its speed of retraction and the amount of released air pressure, followed by some temporally randomly structured low amplitude sound showing some predictable frequency structure related to some moving tongue position within the oral cavity resonator, followed by some temporally structured sound of higher amplitude related to vocal cord vibration, etc. The delay (implying some other expectation) of vocal cord vibration by a few tens of milliseconds has been noted and well documented (Lisker and Abramson, 1964). But in terms of our control model there is no implication that this delay must be specified. Under certain conditions (the ones actually pertaining in this example) silence + burst + vocal cord vibration *means* a delay to vocal cord vibration onset.

In the examples given, two internally structured effects are described. In the one (jaw — tongue interdependence) two articulators have their independent controllability constrained by simple mechanical linkage of some sort. It may be that one articulator is more readily or finely controllable than the other, but the linkage persists giving rise to asymmetrical tongue control to achieve possibly symmetrical tongue positioning about a neutral position. In the other example an aerodynamic effect has probably determined that the vocal cord vibration onset will occur somewhat after the stop's release. Notice that within limits in both these examples the constraint can be overcome: the tongue can be moved up while the jaw is being lowered, the timing of vocal cord vibration onset need not be exactly as determined by the aerodynamic effects. Control of the constraints though, requires at least a. prior knowledge of the rules governing the constraints and b. prediction that the running context is right for this constraint to come into play.

Following Fowler in principle (though perhaps not in detail) the speech production model being described here has a control system and control information which is rather less detailed than has hitherto been proposed. Earlier models called by Fowler 'translation' models, required detailed control from a high, phonological level right down to muscular contraction, thereby implying a high degree of independent controllability of individual articulator movement. We see the same in models of phonology. In distinctive feature theory (Jakobson, Fant and Halle, 1963), for example, very little is claimed regarding the internal relationships within distinctive feature matrices, implying relatively unconstrained marking of + or – in any cell. It is true that there have been some attempts to express constraints existing within a distinctive feature matrix: hence the segment and sequence structure conditions characterising constraints on feature combination within morphemes, usually on a language specific basis but not always (Stanley, 1967).

Gross instructions on their own would often fail to generate a correct articulatory (and hence acoustic) output. The phonetic constraints of the types illustrated above operate automatically to specify detail of articulator movement. It is noticeable, as exemplified, that very often languages vary the structure-specified articulation. In such cases the model allows for a conceptually separate control signal to adjust the low level constraint. This notion does correspond somewhat to Fowler's 'tuning' of co-ordinative structures, though I am taking a less mechanistic approach in general. An adjustment control signal could only be correctly generated, it is emphasized, if the device had drawn on information in the knowledge base

describing in a predictive fashion the nature of the constraint, the context of its occurrence and the limits of its controllability.

It is not enough, though, to establish in the model a means of generating articulations by gross control and a means of tuning those articulations by detailed, knowledge dependent fine control. It is necessary to state the conditions under which such adjustment occurs. In a paper (Morton and Tatham, 1980) the device (then called *Production Instructions*) was described which generates this fine control. It is sensitive to phonological knowledge and to knowledge of the expected perceptual response to an output without fine control. That is, the device accesses a phonology of the language and a generalised model of perception. In the theory of cognitive phonetics the fine tuning generator is responsible for providing a continuously varying signal which, in the knowledge of all low level articulatory constraints, enhances or limits the intrinsic relationships between elements within the articulatory structures as well as intrinsic relationships between articulatory structures (Tatham and Morton, 1980). The adjustments generated and their 'strengths' are determined by two factors:

- the varying demands of the phonology on critical aspects of the gross phonetic system, and
- the predicted stress loading placed on a perceptual system required to respond to the soundwave to be produced.

A further example may clarify the point: the phonology of English is such that the set of vowels seems asymmetrical with regard to articulatory and perceptual spaces. If we assume a relatively linear gross control for tongue height then the keeping separate of some vowels (e.g. [u] and [o] will need more careful or accurate control than with other vowels). That care seems indeed to be exercised is shown in Peterson and Barney (1952, p. 182) where the variability of acoustic placement for the vowel in 'book' and the vowel in 'bird' is shown to be less than that for many other vowels. In addition the probability of perceptual confusion arises when it is semantically critical to distinguish between two words, say, differing only by such a vowel. The device responsible for generating fine control of articulation is triggered to produce an appropriate degree of fine control overlaid on the usual gross control by accessing knowledge of the detail of the phonological inventory of the language and by a prediction that perceptual confusion may arise if control is inadequate, recognising the semantic loading on this segment on this occasion. Fine control results in some combination of more careful articulation, momentary rate of utterance decrease or perhaps the application of a degree of contrastive stress to alert the perceptual system. Indeed on this latter point of alerting the perceptual system we might imagine some subtle dialogue ensuing between a speaker's perceptual model, the articulatory control and the actual perceptual system to which the utterance is being addressed.

Perception in cognitive phonetics involves what can be modelled as two parallel though interdependent processes somewhat analogous to, though not the same as, the Darwin II automaton for visual pattern recognition described in Reek and Edelman (1984), or as two aspects of the same process which can be characterised separately. Both sub-systems incorporate bottom-up and top-down processing: response to the incoming signal is mediated by the knowledge base. One system responds to gross aspects of the signal and is continuously operative, the other responds as necessary to detail in the signal. The gross *vs*. fine distinction being made reflects the two aspects of production discussed earlier.

Perception is modelled as a network (see below) representing the knowledge base, and activation of that network according to certain constraints. Recognition is response to aspects of dynamic patterning within the system, and that response is signalled by convergence of activation patterns with patterns stored in the knowledge base. Since the knowledge base is the network, then convergence means the activation settling to indexed threshold values. An outline of the representation is discussed later, but recognition can be thought of as the spreading of the activation within a part of the overall network to another separately identifiable part. So, in symbolic terms borrowed from the more familiar phonology, the activation pattern of features [+consonantal, +stop, +bilabial, –voice, etc.] permits activation

of some node or pattern of nodes and connections elsewhere in the network appropriately symbolically represented as /p/.

This oversimplification highlights a distinction between perception and reporting of perception by repeating back. Indeed saying 'I heard the sound [p]' seems to me to be a very strange and special thing to do. However, it is true that this does occur and is often the kind of thing we want automatic speech recognition systems to do. In addition, of course, we can successfully mimic sounds and accents we have just heard but are otherwise unfamiliar with. These two pieces of informal data (especially the latter) lend themselves to explanation by postulating a. that the network activated in the perceptual process is the production network, and b. that activation takes time to decay in some sense — meaning that the network must be able to learn and remember.

SUMMARY OF SPEECH PRODUCTION

Speech production is a top-down and bottom-up process. For descriptive purposes the knowledge base is separable into several components. The traditional cognitively based phonology requires a knowledge base including a list of language specific symbolic representations of articulations or sounds. This list has been filtered by the application of explicit constraints from the set of all such sounds usable in language which in turn has been filtered from the set of all sounds which can be made by the human vocal apparatus. The low level constraints detailed in the knowledge base include statements as to what is logically possible, actually possible from the point of view of motor control, perceivable, etc. The high level constraints involve decisions as to what shall be the set of sounds used in this language, given a larger set of potentially usable sounds.

In addition constraints from a low level are set on the rule governed manipulation of the symbols by what is and is not possible for the combination of the selected segments and from a high level by what from this set is preferred by the language community.

In the phonological stages of the encoding of a concept the knowledge base is drawn on to provide an appropriate unique representation of a given word in terms of phonological symbols, and appropriate rules are applied to derive a symbolic representation suitable for proceeding with articulator control.

At the phonetic level the knowledge base contains information as to the appropriate neuromuscular specification of what linguists symbolically represent as phonological units, where these are not so much objects as occurring representations of patterns of excitation and inhibition within some network representation. The neuromuscular specification is in terms of articulatory structures analogous to those described by Fowler. Accessing the knowledge base and sensitivity to what is found there enables the emergence of appropriate gross control signals. Concurrent activation through a knowledge base characterising perception highlights areas of perceptual difficulty in interpreting the acoustic waveform which will, it is predicted, match the gross control signals computed, while access to semantic information highlights areas of articulation to be brought into focus — candidate areas for cognitively generated overlays on the gross control. Additional control signals are generated to supplement the original ones to provide appropriate enhancement to the articulatory effects of the gross system. As may be seen later, although the knowledge base *is* the network it lends itself to separate abstract representation.

SUMMARY OF PERCEPTION

Patterns of excitation and inhibition within a network triggered by the input acoustic signal excite nodes and inter-nodal connections to beyond critical thresholds ensuring the left–to–right successive triggering of higher level networks. The networks themselves are primed (i.e. thresholds are set) in according with the speech production knowledge (that is, their pattern of priming *is* the knowledge of speech production). No mechanistic segmentation of the incoming signal, nor time normalisation is necessary in such a model. Nodes and patterns of excitation become the symbolic representations in a continuous bottom-up and top-down

process like that of COHORT and TRACE (Marslen-Wilson and Welsh, 1978; Elman and McClelland, 1984).

A model of this kind permits explanation of the behaviour observed in a human being asked to repeat the utterance he just heard, portions of that utterance, or respond to questions like 'What was the sound at the end of the word I just said?'. Depending on what the listener is required to do, responses can be generated which draw on the excited perception / production network and which lend themselves to the *incorrect* earlier conclusions that somehow sentences, phrases, words, phonemes have been 'extracted' from the acoustic our listener hears and that he has separately run through a production routine to respond.

In terms of the cognitive mechanisms involved and in terms of the running of these mechanisms, speech production and perception are not different. Human speech behaviour can be seen to involve too many considerations of perception to be adequately modelled as a distinct system. Similarly the spoken responses which can be elicited from a listener to questioning about what he has just heard reveal so much detail of the production of the heard utterance that is seems inadequate to model the perceptual process as distinct from production. The linguist's symbolic representations of what is going on no more lead to postulating a symbolic 'level' within the speaker / hearer than they lead to postulating distinct mechanisms. The symbolic representation is just that — a representation.

THE REPRESENTATION

In the cognitive phonetics model the knowledge base is represented within a multidimensional network of interconnected nodes. The knowledge is the network. The dynamic running model is a spreading activation of the system with the spreading governed by a. the nature of the nodes and b. thresholds set to constrain connection activation. Activation proceeds with variable strength depending on local nodal conditions within the network.

It is not possible here to present more than a tiny fragment of such a network nor to give more than a hint as to the simplest possible implementation of the nodes and their connections. As a first approximation a node can be imagined as some constrained finite automaton, and connections as marked with some appropriate threshold index. Constraints governing node operation and connection threshold are variable dependent on the network's long term and short term 'experience' of earlier activation — that is, the network adapts and can learn. In some sense the knowledge base (an abstraction, of course) is self-adapting within the dynamic system.

Connections between nodes are of two types: possible (= exist) and impossible (= cannot currently exist). Those which do exist have the potential of being activated in a range exemplified here as being from –1 though 0 to +1, where negative numbers indicate inhibition and positive numbers indicate excitation. 0 means unspecified or some balance between excitation and inhibition. Thresholds are set as indices. Thus a inhibitory threshold of –0.3 or an excitatory threshold of +0.3 might be set, preventing spreading of activation until such thresholds are passed. The concept of threshold on connections provides, in effect, a layer of constraint on the associated nodes.

For the purposes of conceptualisation or computation the entire dynamic network can be regarded as existing in a series of time slices, each of which provides a window on the progress of its activation at any one moment.

Within this very general framework, and in an elementary fashion which begs many questions let us consider some examples, proceeding from generalized fragments to fragments which show some recoding of simple existing concepts in phonology / phonetics using familiar terminology for the purposes of illustration.

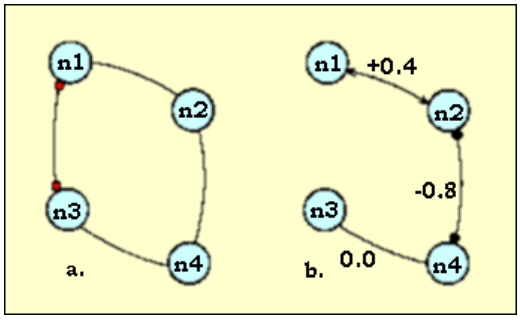

Fig. 1. Generalised network, showing four types of connection with numerically specified thresholds. A simple, unterminated edge represents an unspecified connection; an arrow terminated edge an excited connection; a black circle terminated edge an inhibited connection; a red circle terminated edge an impossible connection.

Figure 1 shows part of a generalized network. There are four nodes: n1, n2, n3 and n4. In a. the two main types of inter-node connection are shown: n1 to n2, n2 to n4, n4 to n3 are possible connections, but n1 to n3 is shown as currently impossible. In b. the possible connections are shown with indices setting threshold levels, with connection n4 to n3 unspecified.

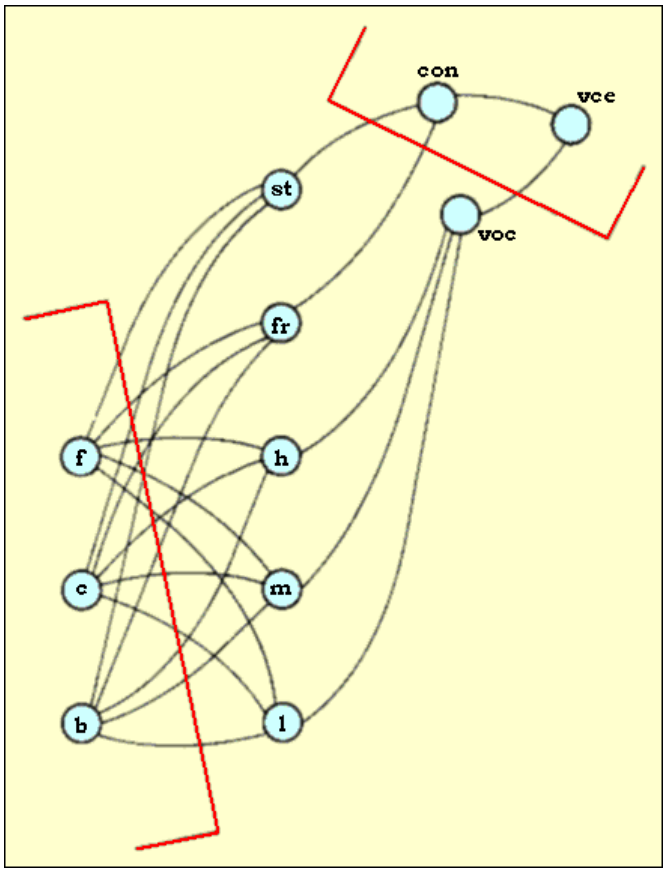

Fig. 2a. Fragment of the phonological knowledge base. Three planes are shown. Abbreviations used in Fig. 2: cons = consonantal, voc = vocalic, vce = voice, st = stop, fr = fricative, h = high, l = low, f = front, c = centre, b = back.

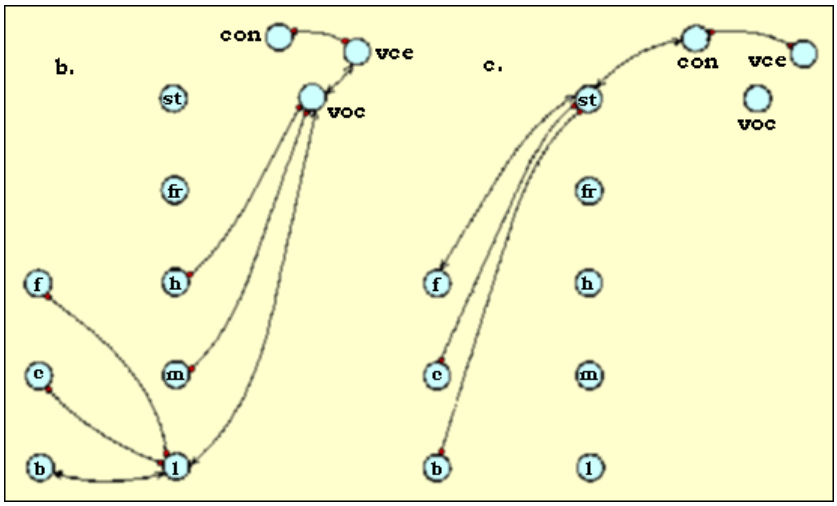

Fig. 2b. Excitation and inhibition connections shown to match the symbolic representation /a/. Fig. 2c. Excitation and inhibition connections shown to match the symbolic representation /t/.

Figure 2 shows a small fragment of that part of the knowledge base roughly corresponding to what we can identify as part of phonology. For illustrative purposes three planes within the phonological knowledge space are shown, each accommodating some nodes with interconnections between nodes on each plane and between planes. Impossible connections are not shown for the sake of clarity, but an example might lie between the node [front] and the node [central]. Figure 2a is what the unspecified network looks like: all connections are unspecified or equally possible, and this roughly corresponds to the distinctive feature representation shown in Table Ia in which any combination of [+/–*feature*] is implied. Figure 2b is the same network with connections set for an activation pattern which might have, in some other plane, the matching (but not derived in any literal sense) symbolic representation /a/, and Fig. 2c shows activation for a pattern matching the symbolic representation /t/. The corresponding distinctive feature representation appears in Table Ib and Ic, respectively. Notice that the two representations are not notationally equivalent. The network representation is far more structured than the distinctive feature representation in as much as hierarchical precedence and internodal (inter-feature in this case) relationships are specified in the network. Just as in a sense in distinctive feature theory the patterns in the matrix *are* the symbolic representation of /a/ and /t/, so they are in this network representation.

	a. S1	S2	S3	S4	S5	…..	S6		b. /a/	c. /t/
Consonantal	+/–	+/–	+/–	+/–	+/–	…..	+/–		–	+
Vocalic	+/–	+/–	+/–	+/–	+/–	…..	+/–		+	–
Stop	+/–	+/–	+/–	+/–	+/–	…..	+/–		–	+
Fricative	+/–	+/–	+/–	+/–	+/–	…..	+/–		–	–
High	+/–	+/–	+/–	+/–	+/–	…..	+/–		–	–
Mid	+/–	+/–	+/–	+/–	+/–	…..	+/–		–	–
Low	+/–	+/–	+/–	+/–	+/–	…..	+/–		+	–
Front	+/–	+/–	+/–	+/–	+/–	…..	+/–		–	–
Centre	+/–	+/–	+/–	+/–	+/–	…..	+/–		–	–
Back	+/–	+/–	+/–	+/–	+/–	…..	+/–		+	–
Voice — etc.	+/–	+/–	+/–	+/–	+/–	…..	+/–		+	–

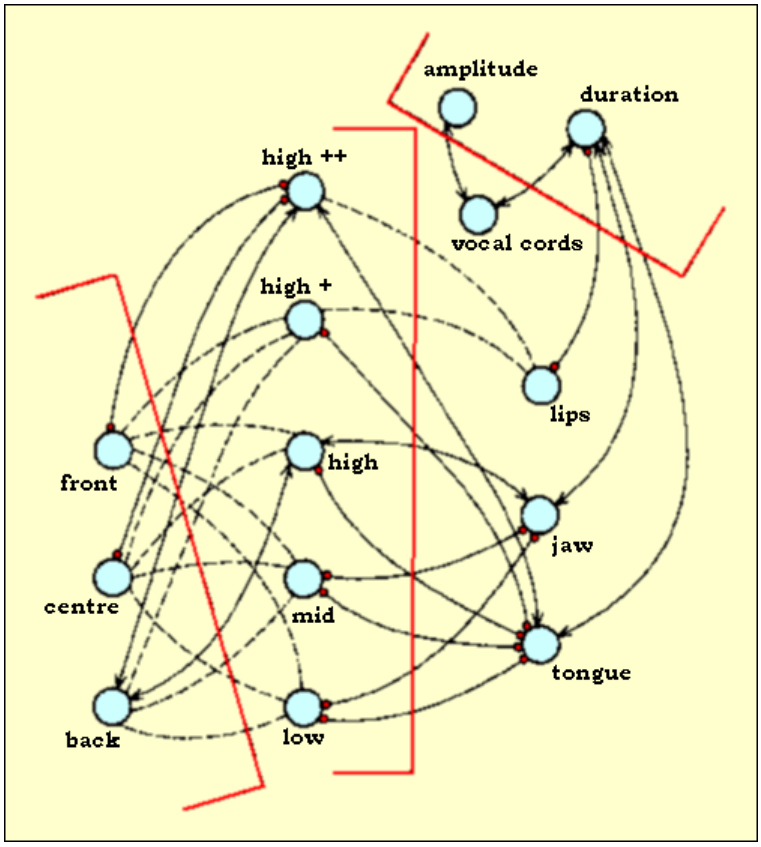

Fig. 3. Fragment of the phonetic knowledge base. Four planes showing excitation and inhibitory connections for [g] — no thresholds marked.

Figure 3 is a fragment of the phonetic knowledge base. Unspecified inter-nodal connections and the connections appropriate for matching the symbolic representation 'intervocalic [g]' in English are shown.

It must be stressed again that the labelling of nodes and the specification of connections here is purely illustrative of the type of network involved in the representation of the knowledge base postulated for speech production and perception. In the dynamic process nodes will be activated in some spreading process according to the knowledge embodied in these specifications, the spreading proceeding from different co-ordinates depending on whether the process is identifiable at any one moment as production or perception. In either case some symbolic representation matches the dynamic network activation pattern as appropriate.

CONCLUSION

In this paper I have presented an outline of a cognitively oriented theory of phonetics drawing attention specifically to the modelling of a speaker/hearer's phonological and phonetic knowledge base. The knowledge base is an abstraction, for theoretical purposes, of the potential of some production/perception system. The system itself is modelled as some dynamic spreading activation network with certain properties. Attention has been drawn at least to the nature of nodes and nodal constraints, including the idea of maxima and minima (–1 to 0 to +1) for activation levels in connections between nodes. In the dynamic model corresponding to (and not separate from — except as a theoretical abstraction) the knowledge base, direction of spreading activation determines whether the network is producing or perceiving, or indeed doing both simultaneously. Thresholds on inter-nodal activation are self-adjusting, providing for a variable knowledge base. Likewise totally inhibited inter-nodal connections might be 'opened up' as part of some learning process. In these senses the dynamic system itself contributes to the evolving knowledge base.

The cognitive theory of phonetics is simpler than most current theories of speech production and perception. Firstly production and perception are collapsed inasmuch as the act of production involves some perceptual prediction and the act of perception involves some production prediction. The knowledge base is therefore shared. Secondly total downward translation of all information in the production model, or upward translation of all information in the perception model is unnecessary, indeed explicitly denied. There is a bottom-up contribution in production and a top-down contribution in perception and two sub-systems are involved.

It is suggested that the implementation of this model as a composite speech synthesis/ automatic speech recognition system (see Bridle (1984, 1985) for ideas about incorporating synthesis into recognition) is a viable proposition which hopefully will overcome many of the problems hitherto encountered in practical speech synthesis and automatic speech recognition systems which have not been conceived as interdependent.

References

Bridle, 1.5. (1984). Challenges and opportunities in techniques for speech pattern processing. In *Proceedings of the 1st international conference on speech technology*, (J. N. Holmes, editor), 191 — 196. Bedford: IFS (Publications) Ltd. Amsterdam: North Holland.

Bridle, J. S. (1985). An approach to speech recognition using synthesis by rule. *In Computer Speech Processing* (F. Fallside and W. A. Woods, editors), in press.

Bridle, J. S., Brown, M. D. and Chamberlain, R. M. (1981). An algorithm for connected word recognition. In *Automatic speech analysis and recognition* (J. — P. Haton, editor), 191 — 204. Dordrecht, Holland: D. Reidel.

Cohen, P. S. and Mercer, R. L. (1975). The phonological component of an automatic speech — recognition system. In *Speech recognition* (D. Raj Reddy, editor), 275 — 319. New York: Academic Press.

Elman, 1. L. and McClelland, J. L. (1984). Speech perception as a cognitive process: the interactive activation model. In *Speech and language: advances in basic research and practice* 10 (N. Lass, editor), pp. 337 — 374. Orlando, Florida: Academic Press.

Erman, L. D. and Lesser, U. R. (1980). The Hearsay — II speech understanding system: A tutorial. In *Trends in speech recognition* (W. A. Lea, editor). 361 — 381. Englewood Cliffs, N.J.: Prentice — Hall.

Fowler, C., Rubin, P., Remez, R. and Turvey, M. T. (1980). Implications for speech production of a general theory of action. In *Language production* 1 (B. Butterworth, editor), London: Academic Press.

Fowler, C. A. (1983). Realism and unrealism: a reply, *Journal of Phonetics*, 11, 303 — 322.

Fowler, C. A. (1985). Current perspectives on language and speech: a critical overview. *In Speech science: recent advances* (R. G. Daniloff, editor), pp. 194 — 278. San Diego: College — Hill Press.

Hammarberg, R. (1982). On redefining coarticulation, *Journal of Phonetics*, 10, 123 — 137.

Holmes, J. N., Mattingly, I. G. and Shearme, J. N. (1964). Speech synthesis by rule, Language and Speech, 7, 127 — 143.

Jakobson, R., Fant, C. G. M. and Halle, M. (1963). *Preliminaries to speech analysis*. Cambridge, MA: M.I.T. Press.

Kelly, J. L. and Gerstman, L. J. (1961). An artificial talker driven from a phonetic input*, Journal of the Acoustical Society of America*, 33, 835 (Abstract).

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. Chicago: University of Chicago Press.

Lisker, L. and Abramson, A. (1964). A cross — language study of voicing in initial stops: acoustical measurements, *Word*, 20, 384 — 422.

Lowerre, B. and Reddy, R. (1980). The Harpy speech understanding system. *In Trends in Speech Recognition.* (W. A. Lea, editor), 340 — 360. Englewood Cliffs, NJ: Prentice — Hall.

Marslen — Wilson, W. D. and Welsh, A. (1978) Processing interactions and lexical access during word recognition in continuous speech, *Cognitive Psychology*, 10, 29 — 63.

Mattingly, I. (1970). Speech synthesis for phonetic and phonological models. In *Status report on speech research SR — 23*. Newhaven: Haskins Laboratories.

Moore, R. K. (1984) Overview of speech input. *In Proceedings of the 1st international conference on speech technology* (J. N. Holmes, editor), 25 — 38. Bedford: IFS (Publications) Ltd. Amsterdam: North Holland.

Morton, K. (1985). Cognitive phonetics — some of the evidence. In a volume edited by R. Channon and L. Shockey, Dordrecht: Foris, *in press*.

Morton, K. and Tatham, M. (1970). The phonetic component. *Occasional papers* 8. University of Essex: Department of Language and Linguistics.

Morton, K. and Tatham, M. (1980) Production instructions. *Occasional papers* 23. University of Essex: Department of Language and Linguistics.

Paliwal, K. K. and Ainsworth, W. A. (1985). Dynamic frequency warping for speaker adaptation in automatic speech recognition, *Journal of Phonetics*, 13, 123 — 134.

Parker, F. and Walsh, T. (1985). Mentalism vs. physicalism: a comment on Hammarberg and Fowler, *Journal of Phonetics*, 13, 147 — 153.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of the identification of vowels, *Journal of the Acoustical Society of America*, 24, 175 — 184.

Reeke, 6. N. Jr. and Edelman, 6. M. (1984). Selective networks and recognition automata. In *Computer culture* (H. R. Pagels, editor), 181 — 201. New York*: Annals of the New York Academy of Sciences* 426.

Sakoe, H. and Chiba, 5. (1978). Dynamic programming algorithm for spoken word recognition, *IEEE Trans. Acoust. Speech Signal Process.*, 26, 43 — 49.

Stanley, R. (1967). Redundancy rules in phonology, *Language*, 43, 393 — 436.

Tatham, M. (1970a). Coarticulation and phonetic competence. *Occasional papers* 8. University of Essex: Department of Language and Linguistics.

Tatham, M. (1970b). Articulatory speech synthesis by rule: implementation of a theory of speech production. *Working papers in linguistics*. The Ohio State University: Computer and Information Science Research Center.

Tatham, M. (1971). Classifying allophones, *Language and Speech*, 14, 140 — 145.

Tatham, M. (1984). Towards a cognitive phonetics, *Journal of Phonetics*, 12, 37 — 47.

Tatham, M. Daniloff, G. 6. and Hoffman, P. R. (1985). Electromyographic invariance of lip closure for /p/ — /b/. In *Speech science: recent advances*, (R. G. Daniloff, editor), pp.279 — 310. San Diego: College Hill Press.

Tatham, M. and Morton, K. (1980). Precision. *Occasional Papers* 23. University of Essex: Department of Language and Linguistics.

Witten, I. H. (1982). *Principles of computer speech*. London: Academic Press.