PRAGMATIC PHONETICS

Katherine Morton

Reproduced from: Morton, K. (1992) Pragmatic Phonetics. In W.A. Ainsworth (ed.) *Advances in Speech, Hearing and Language Processing*, Vol. 2, pp. 17-53. London JAI Press

INTRODUCTION

Phonetics is usually considered to be concerned with the physical processes involved in speech production and perception. Models of speech production deal with the motor, aerodynamic and acoustic processes involved in producing sound signals which are sufficient to trigger perception patterns in the listener, and which correlate with the idea the speaker intends to communicate. In the study of language production, phonetics deals with the physical aspects of sound production, but the way the patterning of speech sounds is organized is dealt with by phonology.

Phonology therefore is seen as the linguistic component which describes the meaningful sound patterns which trigger physical phonetic processes within a language. Although it is incorrect to regard most modern linguistic descriptions as prescriptive or algorithmic in nature, this interpretation often occurs when phonology and phonetics are applied to speech synthesis and speech recognition systems. Some difficulties have resulted from this interpretation and are discussed in this paper.

The paper is divided into two parts: Part I is concerned with explaining the need for going beyond some of the constraints imposed by the metatheory of traditional theoretical linguistics. This section sketches part of a model that might go beyond these constraints in the restricted area of converting abstract intonation patterns into continuous movements of fundamental frequency in a set of utterances.

Part II outlines a way to generate natural fundamental frequency contours in synthetic speech systems without the need for a rule set. Synthesis is used here as a test bed for changes to the model suggested in Part I. A method of normalizing data gathered from human speech by using neural networks is also presented.

PART I: BEYOND TRADITIONAL LINGUISTICS

2. SENTENCES AND UTTERANCES

In theoretical linguistics, the object called sentence defines the domain over which some thought or idea is expressed. A sentence is an abstraction which can exist without reference to other sentences. The meaning of a particular sentence can be taken in isolation and need not be dependent on the meaning of adjacent sentences. However, there are many cases of sentence production where it is possible to define a domain over which a thought is expressed in which meaning varies depending on a number of contextual relationships. This object can be thought of as an actualization or realization of a sentence. It is a description of something which exists in the real world of human language and is commonly called an utterance (Lyons, 1981).

As with the term sentence in linguistics, the term utterance is quite often used with systematic ambiguity: ‘utterance’ can mean a piece of actual speech (i.e. a real world object) or can refer to the description of actual speech (an abstract object). In this paper, the term ‘utterance’ means an abstract object describing real speech and one which can be derived by reference to its related sentence and the semantic and pragmatic context of that sentence.

One feature of sentences that can formally be accounted for in theoretical linguistics is surface structure ambiguity as shown in the classic example ‘Flying planes can be dangerous’, or in a more common example ‘She is an English teacher.’ Utterances of the sentence show this ambiguity less frequently than written sentences which suggests that the semantic and pragmatic context of the collection of utterances taken together can lead to only one interpretation. The listener may not even consider that other interpretations of the underlying sentence are possible.

In those cases where the listener is not aware of alternative readings, it is not clear whether the listener himself has disambiguated on the basis of an interpretation of the utterance. It is just as possible that the speaker has removed the potential ambiguities in the utterances which the sentence might carry. He could do this by making changes to the selection of the phonological features of the specific sentence. For example, such changes may occur in the prosodics, with adjustments made to rhythm, stress, and intonation.

Although in theoretical linguistics disambiguation of this type (non-syntactic) can be ascribed to the phonological component, that explanation is not within the spirit of linguistics itself. This would require phonology to interpret the sentence by providing several parallel and equally weighted phonological outputs and in addition, to select which alternative is the most suitable. Although a set of descriptions of alternative outputs are possible, there are no grounds in the theory for making this kind of choice. That is, phonology can specify different abstract patterns associated with speaking ‘flying planes can be dangerous’ but not choose which is the most suitable as a utterance in context, nor how to speak it.

Speakers do not normally speak multiple versions of sentences; we may sometimes be aware of the different versions indicated by phonology, and choose one of these. But the question of modelling a selection mechanism does not arise in theoretical linguistics. However a full description for speech requires inputs from sources other than the underlying sentence which would provide the additional information needed as the basis for choice.

These additional inputs contain information concerning semantic context and meanings of adjacent sentences. There also should be input from a subcomponent which specifies attitudes, moods and beliefs of the speaker. The result of choosing among phonological alternatives together with contextual information is virtually always a single utterance.

3. LIMITS OF PHONOLOGY

Assignment of a phonetic interpretation to a sentence traditionally falls within the scope of phonology (i.e. a specification for subsequent phonetic realization). Stated simply, phonology characterizes the processes from which a selection of interpretations are drawn which might operate on an incoming sentence. The output of phonology consists of a string of phonetic objects specified with sufficient information to enable a phonetic component to characterize the changes which occur in linking the output of phonology to an acoustic output through a set of phonetic rules.

The phonetic component in the standard model is essentially automatic and assigns no new information that is linguistically relevant. And in the case where a sentence might have several phonological interpretations, this model would characterize several corresponding phonetic realizations. Again, there is no way of choosing which is the most suitable for each utterance.

Phonology provides ideal segmental representations of sentences; these abstract representations can be seen to underlie utterances. Prosodic features of sentences are also specified in phonology and it is this area that is most likely to provide variant interpretations associated with ambiguity. For example: ‘She is an English teacher’ with emphasis on ‘English’ compared with ‘She is an English teacher’ with emphasis on ‘teacher’. These utterances may be disambiguated by change in stress, or intonation. Accounting for the knowledge which allows the speaker to produce such utterances does not fall within traditional linguistic metatheory. The linguist’s task stops at the end of phonology; reinterpretation of phonological output in order to produce acceptable utterances is therefore constrained by the objectives of the theory.

However, a model which could deal with selection from among several possibilities would be of use to researchers in speech. For example, in building a speech synthesis system, the synthesizer must speak one utterance rather than a series of alternatives. There are applications in other areas such as speech recognition, sociolinguistics, language pathology, dictionary pronunciation rules, language learning, where a model which can specify selection among alternatives would be necessary.

Both theoretical and applied research in these fields of study could find it useful to be able to choose between alternative readings of sentences due to ambiguous syntactic output, and to be able to deal with more subtle alternatives concerned with conveying an indication of the speaker’s attitude, mood or intent. This nuance cannot usually be decoded from plain text, i.e. in the surface specification of a sentence at the end of the syntactic component.

4. ADDING TO THE LINGUISTIC MODEL

For the purpose of selecting between sentence readings, for incorporating pragmatic constraints, for characterizing beyond the sentence domain in general, we need to add to the linguistic model. Ideally, a complete theory of language would be extended to include all aspects of encoding thought as sound waves, and of constructing algorithms for production, in addition to the usual descriptions and characterizations of knowledge. Incorporating a mechanism for including pragmatic and contextual aspects of language and speech production should be developed. There is no alternative currently available which is as comprehensive, or formal as Chomskyan sentence grammar and its derivatives.

I intend in this paper to propose an interim way of dealing with the lack of an alternative to traditional linguistics that has confined itself mainly to sentence characterization. This is not an attack nor is it a formal proposal to extend or replace part of the existing model.

Following on from Tatham (1984, 1990), in which a cognitive theory of phonetics is proposed, I would like to suggest a component called the ‘Selectional Subcomponent’ which could stand between phonology and phonetics within a Cognitive Phonetics. It would provide a means of selecting a single alternative from a set of phonological representations of a single sentence, and a way of incorporating the prosodic variants which cue the listener to perceive the attitude, mood, etc of the speaker.

Cognitive phonetics suggests that there are cognitive processes involved in the realization of articulation which are not accounted for in physical phonetics or in phonological theory. These processes usually involve choosing to impose constraints on physical processes in order to facilitate perception. The Selectional Subcomponent I describe below constrains phonetics to realize only one of the set of alternative pronunciations of a sentence characterized by phonology. Like cognitive phonetics the component has additional inputs in addition to those from phonology.

5. THE SELECTIONAL SUBCOMPONENT

There are several questions to answer about the identity of such a component:

1. where does it stand in relation to theoretical linguistics?
2. what would be its input and output?
3. what is its internal structure?
4. would it encroach on any area currently satisfactorily covered in traditional linguistics?

The interim model being proposed involves the derivation of an utterance from a sentence at the speech production level. It does not include alternative syntactic solutions which may depend on context, but only with varying the way in which a particular syntactic sentence might be spoken depending on such a context. That is, given a single sentence of a particular syntactic and lexical form, phonology can provide alternative phonetic interpretations of that sentence. The sentence characterization provided by theoretical phonology is taken as the entry point to the selectional component which is different in nature from the other linguistic components in one major respect: it contains algorithms for speaking the appropriate version of the sentence. Figure 1 illustrates where the selectional subcomponent stands in relation to other components.

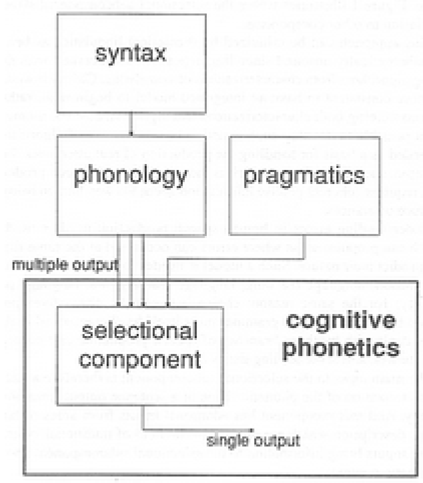


Figure 1. The relationship between the proposed selectional component and other components of the grammar. The purpose of the component is to filter the various alternative phonetic interpretations of a sentence provided by phonology to output just one utterance for phonetic realization.

This approach can be criticized by theoretical linguistics as being metatheoretically unsound since linguistics is not concerned with deriving algorithms from characterizations of knowledge. Clearly it would be more consistent to have an integrated model to begin with, rather than combining both characterizations and algorithms. At the moment it not possible to develop such a model; however, a set of algorithms is needed as a basis for handling the production of real utterances. For example, speech synthesis, which is the simulation of speech production, requires coherent precise information about the way human beings produce utterances.

Understanding errors in human speech production needs a model which can pinpoint areas where errors can occur and at the same time can predict their nature. Such a model is needed in the areas of speech recognition, language teaching, language learning, and language pathology; for the same reason: choice between the alternatives permitted by the rules of the grammar must itself be characterized in the areas dealt with by these branches of speech production and perception studies, and sentence linguistics does not do this.

The main input to the selectional subcomponent is therefore a static characterization of the phonetic shape of a sentence output from phonology. And this component has additional inputs from areas of language description which are not within the field of traditional syntax. These inputs bring information to the selectional subcomponent about pragmatic context.

Context is a vague term, but is generally taken to mean factors which influence the form and appropriate specification of an utterance (Levinson, 1983). They include environmental factors such as

- the social situation in which speech production occurs,
- the psychological state of speaker and listener,
- the social relationship between speaker and listener.

Linguistic factors are also important such as how adjacent sentences are related to the thoughts the speaker is communicating. The domain ‘context’ is wide and a specification of context is potentially capable of considerably altering an utterance.

In Part II of this paper, I shall consider only one aspect of context: the pragmatic context of the speaker’s mood as affected by what he or she says and hears, and which is intended to be communicated to the listener. A listener can usually immediately tell if a speaker is pleased, irritated, surprised, angry, frightened, etc both as a general mood or as a result of the speaker receiving information from him or her.. The actual words used in the utterance need not explicitly convey the mood. In fact, written words usually communicate a plain message only. Thus the manner of speech can be thought of as overlaying the plain message with some kind of additional comment. Or in other words, this comment can be considered as information added to the basic message or information which modulates the message in a particular way.

The examples in Part II illustrate how a speaker’s mood can transform the intonation and internal duration of a neutral version of the utterance. A neutral version of a sentence constitutes the linguistic input to the selectional subcomponent; it is abstract, the sound pattern is described by the phonological component, and no context of any kind is specified.

6. INTONATION

The intonation pattern of a sentence is abstract and specified as an invariant pattern characterized by the intonation rules within the phonological component. The task is to take this pattern and turn it into a specification for the time-governed changes that occur in the rate of the speaker’s vocal cord vibrations while speaking the sentence as an utterance. In order to show how this might be done, it was first necessary to build a model which could carry out the task; the second step was to show what is needed while the model is running.

Since intonation is correlated with changes in fundamental frequency of the speech waveform, it was necessary to assemble data about the actual f0 contours of selected spoken sentences. Unfortunately, the literature was not particularly helpful in this case because the phonological descriptive system used to describe intonation is abstract. Thus the observation that within the normal rise-fall intonation contour of an utterance there could occur a high fall on the stressed syllable of a word which contrasted with a previous word, was only useful to indicate what to look for when examining a real utterance. The observation could not be turned into a prescription for simulating the effect.

Looking at speech data shows immediately that these descriptions are simplified; they highlight essential areas of intonation pattern changes but they do not attempt to describe all the changes that need to be simulated. Therefore it was necessary to gather the data from a speaker; he was presented with neutral sentences and asked to speak these sentences in different mood styles overlaid on neutral speech, e.g. happy, surprised, gloomy, etc. This experiment is detailed below in Part II.

7. THE INTERNAL FORM OF THE MODEL

The selectional subcomponent being modelled needs a formal characterization of its output, a set of rules linking this output with the formal description of the phonetic shape of the sentence intonation that constitutes the neutral input. This requirement comes from a particular tradition in linguistics. The following is a summary of the background linguistic model used which characterizes the sentence, and a description of automatic physical phonetics, along with a suggestion as to how the proposed selectional subcomponent could provide a useful link between them.

In order to model language, linguistics draws up rule sets which characterize what the speaker knows of the semantic, syntactic and phonological structure of his language. The rules operate on elements such as syntactic categories or phonological segments. The rules of the language are said to be known to the speaker and constitute part of what is called linguistic competence. In this approach to the study of language there is no attempt to say how these rules are actually represented in the speaker’s mind although it is implied that there might be some kind of representation which could be related to mental constructs. For practical purposes, the rules provide a formalism for setting out hypothesized representations, and form knowledge bases which could be drawn upon if an algorithmic model were designed to produce sentences (see Figure 2).

In this model, linguistic descriptions themselves are generalizations about language properties contained within the knowledge bases, but there is as yet no information about how to select which description to use for anyone sentence production. Thus a set of linguistic descriptions is not in an algorithmic or prescriptive form.

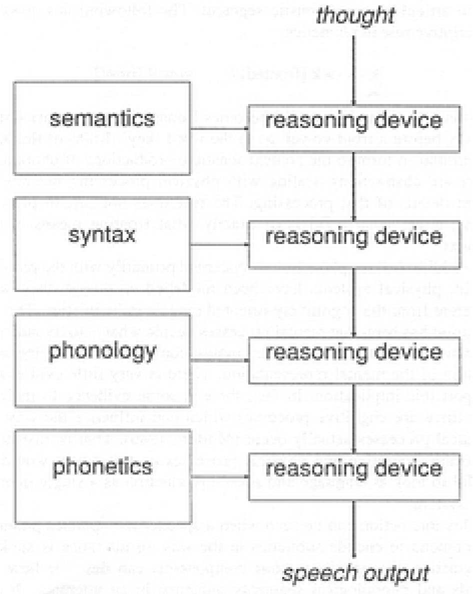


Figure 2. The relationship between the various components of theoretical linguistics and a procedure for producing sound wave encodings of thought.

Phonetics has taken a different, non-mentalistic approach. In contrast with semantics, syntax, and phonology where the material is cognitive in origin, the subject matter of phonetics is mainly physical. Thus phonetics is concerned with modelling the neuro-physiological mechanism of speech production, the aerodynamic system involved, and the resulting acoustic output. Phonetics has only relatively recently used the idea of the model structure of rule set and accessing algorithm (Ladefoged, 1967). For example, the phenomenon of coarticulation is often described in terms of rules transforming the physical specification of an articulatory or acoustic segment. The following is a possible descriptive rule in phonetics

k → k [fronted] / − vowel [front]

That is, 'the neutral target [k] becomes fronted when it occurs immediately before a front vowel' as in the word 'key'. Rules of this kind are similar in form to the context sensitive productions of phonology. They are abstractions dealing with physical processing but are not about details of that processing. The rule does not say, in physical terms, either what a [k] is or exactly what fronting means in this context.

In addition, since phonetics is concerned primarily with the physical world, physical systems have been modelled as though they were different from the cognitively-oriented models in linguistics. The implication has been that mental processes decide what is to be said, this is then translated into a physical representation which is a degraded version of the mental representation. There is very little evidence to support this implication. In fact, there is some evidence to indicate that there are cognitive processes which can influence the way the physical processes actually occur (Morton, 1986). That is, instead of regarding cognitive and physical processes as separate it would be useful to look at language and speech production as a single interactive system.

This interaction can be seen when a speaker manipulates phonetic phenomena to encode subtleties in the way an utterance is spoken. Linguistics, through its various components, can describe how the words and phonological segments sequence in an utterance. It can also describe in general terms the observation in pragmatics (Levinson 1983) that variations of prosodics will trigger subtle perceptions of a speaker's intentions that go beyond the plain message. But it cannot tell us in a detailed way just what events contribute to the perception of emotion or other overtones.

However, it is precisely these subtleties which characterize the naturalness of human speech. This can be seen particularly when human speech is contrasted with synthetic speech. The difference is especially clear when a synthetic utterance generated by rule is compared with the same utterance produced using copy synthesis.[1] Rule generated synthesis lacks naturalness, whereas the copy-generated version, analyzed from a real human utterance often sounds so natural that it is difficult to distinguish it from human speech.

The difference is accounted for by observing that the rules are incomplete in the linguistically-based rule system. It is said that we need to know more about coarticulation, or relative timing of segments, or how prosodics is generated in order to be able to add to the rule set. It is believed that with more completely specified descriptions natural sounding synthetic speech could be produced. This is probably true, but so far refinement of the rules has not brought the expected success.

8. LINKING THE ABSTRACT AND REAL WORLDS

It has been suggested (Tatham, 1986) that a major difficulty in building a complete linguistic-phonetic descriptive system useful for applications work is that there is no way of linking the abstract ideal form of linguistic representations with physical representations. An attempt has been made, grounded in phonetics, to model the interaction between cognitive and physical phenomena in speech production, called ‘cognitive phonetics’. A selectional subcomponent as described in this paper could form part of a cognitively oriented phonetics since it functions by transforming the abstract linguistic output into part of the representation of a speech waveform.

The general problem of satisfactorily providing a link between abstract descriptions of sentences and actual utterances, the lack of a coherent formalization of a suitable approach and constraints on type of rule, lead to a consideration of a different approach from the rule - based model. The connectionist approach of parallel distributed processing might assist in building a device which could

1. evaluate the variability discovered in examples of human utterances which show overall changes of fundamental frequency correlated with pragmatically determined nuances in intonation, and
2. learn to relate abstract phonological descriptions with these evaluations.

In setting out the model, then, the formalism used is that of the neural network in a multi-layered perceptron configuration. This will be described more fully in Part II. As explained this particular formalism was chosen because of the difficulty of establishing an explicit set of production rules to make the link between the abstract phonological characterization of sentences at the output of phonology and the real world specification of the speech waveform associated with the corresponding utterance.

PART II: MODELLING THE NATURALNESS OF SPEECH

9. INTRODUCTION

Part II is concerned with describing how the speech produced by a rule-based speech synthesis system might be improved by adding a module based on implementing the selectional subcomponent outlined in Part I. It describes an attempt to improve naturalness of rule based synthesis using a neural network. The objective is to identify some aspects of what constitutes naturalness in speech and to associate these with abstract linguistic descriptions.

10. SPEECH SYNTHESIS

A speech synthesis system is a good testing ground for models of speech production if those models do not address semantic or syntactic aspects of sentence production, but are concerned only with phonological and phonetic properties of an utterance. A speech synthesizer can be made to sound virtually indistinguishable from a human being if a natural human utterance is carefully analyzed, either by hand or automatically, into synthesizer parameter values which can then be used to drive it. Although the resulting synthetic speech is very good, the analysis is time consuming and the results are replications of previously spoken utterances. Novel utterances cannot be produced, consequently use of such speech is limited.

Good standard text-to-speech synthesis systems do not sound natural for two reasons:

1. conjoining individual sounds has proved difficult, and
2. prosodic features are not rendered adequately.

One reason for poor prosodic rendering is that such effects are only minimally marked in the input text by using punctuation. Therefore prosodic features such as intonation, stress and rhythm have to be derived from generalizations which must apply across all utterances, and which may be inappropriate for the actual utterance being synthesized. The perceptual effect created by the speech is a rather monotonous and machine-like quality. The algorithms used in synthesis by rule for generating prosodic features are necessarily safe. That is, they are designed to avoid extremes of f0 and other changes that might occur if a speaker is expressing surprise or anger.

Current speech synthesis relies on rules available from theoretical linguistics which are not the same as those needed to produce utterances output from the synthesizer. Assuming there are no errors in linguistic rules, it could be the case that in current synthesis systems we are listening to spoken sentences rather than spoken utterances. It seems to me that this is the very reason why it has been a mistake to take descriptive rules of linguistics and apply them to produce utterances.

In fact, linguistic rules are concerned with describing idealizations or abstractions of language rather than of utterances. These rules thus remove variability, which is precisely the feature of spoken language which must be adequately described so that it can be inserted in simulation models. However, variability in speech is difficult to describe. Modelling variability so that it can be reproduced in synthesis has not been successful to date: one reason for this is that gathering the data which can form the basis of a model of variability is difficult. In the next section, I shall describe using a neural network for the purpose of dealing with variability in speech.

11. USING SYNTHETIC SPEECH AS A TEST BED

I have assumed that improving existing rules in even a very good synthesis system will not produce a more natural output for the reasons already mentioned:

a. existing rules are derived from characterizations of sentences rather than utterances, and

b. there is no good model of variability in human speech.

Within existing systems, there is probably still scope for refining the algorithms that calculate transitions between segments or for improving durational relationships between segments, but these improvements would required a great deal of data collection and generalization. It is not clear that enough information could be determined using existing techniques that would incorporate the variability into a suitable speech model. I have taken the approach that improvements in suprasegmental effects can considerably add to increasing the naturalness of the output of synthesis systems.

In the examples reported here, I have assumed the following:

1. That current segmental descriptions are adequate; that is, that the output of the phonological stage of the system constitutes a string of segments which can be conjoined by transition rules, and that the descriptions of each of the phonetic segments used in the synthesis model are good enough to use to produce a base upon which pragmatic features can be overlaid. This assumption is made despite the observation made above that there is still some room for improvement without adding pragmatic features.

2. That pragmatic descriptions are available as to how an utterance is to be spoken, say, angrily or with surprise. These descriptions would be defined in a pragmatics section of the grammar, and be accessible to the selectional subcomponent. They enable context-free sentences to become utterances. For an account of such a pragmatic component and its relationship to syntax and semantics, see Levinson (1983).

12. SETTING UP THE SELECTIONAL MODEL

The research reported here begins by taking the outputs from two components within linguistics. These are (a) a standard description, provided by phonology, of phonetically specified sentences, and (b) pragmatic markers for the same sentences from a pragmatic description. This combination constitutes the input to the selectional component. An example input might be the phonetic specification, expressed as a string of systematic phonetic segments, of the sentence ‘The weather seems to be getting better’, together with the pragmatic marker [with relief].

As explained earlier it seemed very difficult to relate these inputs to how human beings actually manipulate fundamental frequency contours under such conditions. In principle it should be possible to write a set of production rules to express the relationship, but I felt it would be useful to employ the learning properties of neural networks to solve the association problem. Neural networks, in the multi-layered perceptron configuration, are able to establish relationships between paired items by a process of exposure to repeated examples of paired data. In the model described here, one half of the data for the learning process was the phonological description together with the appropriate pragmatic marker.

The other half of the training data consisted of example f0 contours taken from a speaker instructed to add particular emotions to neutral sentences. In setting up a sample model, a speaker, a native of southern England, was asked to produce utterances in whatever way he felt necessary to convey certain emotional overlays he was prompted with. That is, a human being was asked to do precisely what would be required of a speech synthesis system: given an idealized symbolic representation of a sentence and an indication of how it is to be spoken, produce the appropriate speech.

Two sentences only were used in this feasibility study:

‘There’ll be a high tide in two hours.’

‘I think that’s probably true.’

Each sentence was to be spoken in eight different ways:

neutrally
questioningly
to convey happiness
excitement
gloom
disappointment
contrast
surprise

Thus there were 16 possibilities. These were written on cards in sentence form in ordinary orthography along with one of the eight markers to indicate how the sentence was to be spoken. Each of the 16 cards was presented to the speaker 25 times, with the cards randomized after each set of 16 utterances. This made a total of 400 utterances which were recorded using a condenser microphone and a Sony Fl digital tape recorder.

The speaker intended to produce examples of spoken utterances which for him matched what was written on the prompt cards. Informal tests among a number of people listening to the recorded material confirmed that a third party could, on listening, determine not only the meaning of the two sentences but also the emotion the speaker had intended to convey. There was some perceptual confusion between happiness and excitement, and between gloom and disappointment. Since the purpose here is to determine comparative subtleties of speech, an overlap in the data was not surprising. Not all speech is correctly interpreted by listeners. Part of making synthesized speech more realistic involves replicating these effects, and it was interesting to see whether the same confusions would arise later when the utterances were synthesized.

The objective was to build a device to simulate what the human speaker had produced during the recording session, which was to consistently associate a linguistic description (the orthography and pragmatic marker) with his intended acoustic signal. The test of the adequacy of the device, and hence the usefulness of a selectional subcomponent, would be a listening test similar to the one performed on the human speech.

13. THE ASSOCIATION DEVICE

As mentioned earlier, it was decided that the most appropriate way of building a device to model the selectional subcomponent was a neural network. A number of considerations led to this decision:

1. The input to the device is an idealized representation of an intonation pattern, together with pragmatic markers. The output from the device is an actual fundamental frequency curve which could be represented graphically as movement through time of a physical quantity. A neural network is a good way of bridging the gap between abstract and real in circumstances where no methodology exists for doing so. It is particularly useful in the current situation where there is no good metatheoretical basis for some other model.
2. A description of the detail of the f0 curve is extremely difficult to formulate: descriptions in the literature are very informal and cannot account for the detail and variability present in the signal. But in order to simulate human speech with adequate naturalness, the detail has to be reproduced. The method described below captures that detail from the f0 graph, and requires no other representation.
3. Formulating an adequate rule set which describes this detail and which could be related to the underlying abstract representation is currently too difficult. However a mechanism which can learn the associations for itself would implicitly capture the detail needed, since neural networks have the ability to learn the relationship between pairs of descriptions without having to be told what those relationships are. A rule-based model would require explicit formulation of the associations.

14. DATA REDUCTION

One of the problems in phonetics is discriminating between intended detail in the acoustic signal derived directly from the input, and the variability which can be detected when the same utterance is repeated by a speaker; the second type therefore results from lack of precision in the articulatory system. In the case of fundamental frequency curves, an example would be the so-called micro-intonation seen at boundaries between obstruents and vowels, and which derives from the effects of supra-glottal perturbations in airflow and air pressure. The detail of this variability is not usually taken as linguistically significant; it merely shows that speech production is not completely stable. But the detail generated by the higher level variable input to the speech process is significant; this latter is the detail I wished to replicate here.

A neural network is also useful in discriminating between these two types of detail, as will be discussed below. The network appears to reduce the data in these examples in such a way that detail due to the speech production process is preserved, but removes detail due to instability in the system. How this is possible is not clear, though it may be that the network has the effect of a sophisticated averaging technique. If this is the case, use of a network is an advantage, since it appears to deal with the detail not adequately described in the literature.

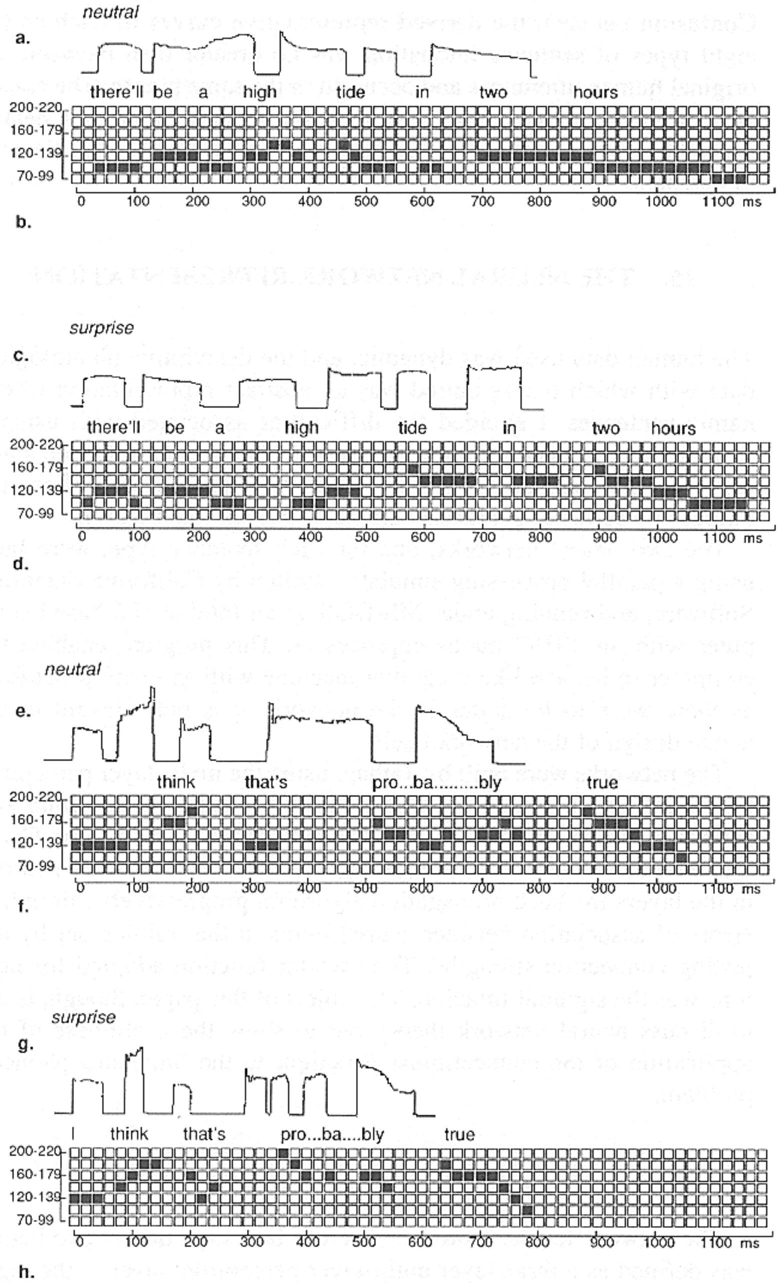


Figure 3. Sample human fundamental frequency tracings, quantized representations presented to the neural network. (a) ‘There’ll be a high tide in two hours’ – spoken neutrally. (b) The network’s reduction of 25 different versions of (a). (c) ‘There’ll be a high tide in two hours’ – spoken with surprise. (d) The network’s reduction of 25 versions of (c). (d), (e), (f), (g) and (h): as for (a), (b), (c) and (d) respectively, but for the utterance ‘I think that’s probably true’.

Figure 3 shows sample fundamental frequency tracings of the human speech, together with the quantized representations presented to the neural network. Representative curves are also shown computed by the neural network after training on 25 versions of each. The human f0 curves were obtained by passing the recording through a Frøkjær-Jensen Pitch Meter set to output a linear display.

This is not a demanding task for the neural network, and the representative curve was arrived at after around 100 training passes through the data. The object of the work, however, was not to optimize the network (which was considerably over-specified) but to develop a method of data reduction for the purposes of removing unwanted detail from the signal, and to present the data in a useful graphical form.

When the derived representative curve was used to synthesize the utterance, the resultant speech was found to be fairly natural, although not as natural as copy synthesis versions of the original 25 utterances. Confusion between the derived representative curves of each of the eight types of sentence intonation was no greater than between the original human utterances and occurred in the same places. The reason why some naturalness was lost may be because removing details caused by instabilities in the human production system was noticeable to listeners.

15. THE NEURAL NETWORK REPRESENTATION

The human data used was dynamic, and the descriptive phonological data with which it was paired was an abstract representation of dynamic sentences. I avoided the difficulties associated with using a dynamic neural network by transforming both sets of data into static graphical representations similar to those often used to show the output of a series of synthesis parameter frames (see below).

The two neural networks, one for each sentence type, were built using a parallel processing simulator written by California Scientific Software, and running under MS-DOS on an Intel 80386-based computer with an 80387 maths coprocessor. This program enabled the computer to behave like a parallel machine with as many processors as there were to be nodes in the network; it is not relevant to the actual design of the network itself.

The networks were built by Tatham using the multi-layer perceptron model and an implementation of the back propagation algorithm as a supervised learning scheme, using the network as an associator. Given initial randomization of the weightings on connections between neurons in the layers the back propagation algorithm progressively minimizes errors of association between paired items in the training set by adjusting connection strengths. The transfer function adopted for neurons was the sigmoid function. The object of this paper, though, is not to discuss neural network theory but to show the usefulness of the application of the connectionist paradigm to the linguistic-phonetic problem.

15.1 *'There'll be a high tide in two hours.'*

The network for the sentence 'There'll be a high tide in two hours' was defined as a three layer multi-layer perceptron: layer 1, the input layer, comprising 112 neurons; layer 2, a single hidden layer, comprising 75 neurons; layer 3, the output layer, comprising 420 neurons. The program established 8475 connections between the input and hidden layers, and 31,920 connections between the hidden and output layers. Weightings were initially randomized.

The input schema for the network was an arrangement of the 112 neurons in a graphical matrix of 4 rows by 28 columns. The matrix was divided into three areas:

1. an area for representing the idealized intonation pattern appropriate to a sentence of this phonological type;
2. an area reserved for indicating the eight pragmatic markers;
3. an area for indicating which of the 25 samples of each of the eight versions was being trained.

1. The intonation pattern was represented on a 2-row by 21-column area of the matrix as follows:

□□□□□□□■□■■■□□□■■■□□□
■■□■■□■□□□□□□■□□□□■■■

Rows indicate pitch level and columns represent progression of the pitch through the sentence (phonology describes sequencing of phonological elements, but does not specify duration or time in general as physical quantities). A dot indicates a non-firing neuron and a = indicates a fully firing neuron. Neurons were set to fire where segments were phonologically [+voice] –

that is, could sensibly have pitch assigned. Thus an abstract curve was established as input to the network, corresponding to the contour:

There'll be a high tide in two hours.

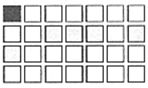
used in more traditional descriptions.

The phonological description of intonation provides a notional representation of perceived change in pitch during the sentence. It can adequately indicate the general direction of change of pitch. This abstract pattern, characterizing a neutral version in phonology, was used as the basis for all eight versions of the sentence. The network would know which of the eight variants of the sentences was being learned because pragmatic markers were given to it along with the canonical intonation curve.

(2) Pragmatic markers were indicated on an area of the matrix consisting of one row and 21 columns. For indicating a neutral version of the sentence the leftmost two neurons were fired, for the happy version the next two were fired, for excited the next two, and so on, thus:

NNHHEEGGDDCCOOQQ

(3) Identifying which of the 25 samples for each version of the sentence was being trained was shown on a 4-row by 7-column area of the matrix. For sample 1, the top left neuron was fired. This is labelled position [1,1]; that is, the neuron in the first column at row 1. For sample 2, the neuron in position [1,2] was fired (column 1, row 2); for sample 3 position [1,3], ... for sample 13, position [4,1], ... for sample 20, position [5,4], ... etc. Thus:



These three areas were combined into the 4-row by 28-column matrix as shown in Figure 4. Figure 5 shows two sample input matrices.

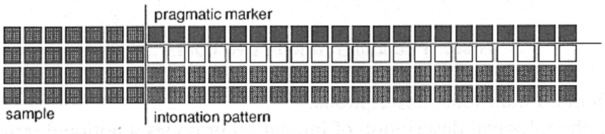


Figure 4. The input matrix of 4 columns by 28 rows showing the areas designated to indicate (a) sample numbers (b) pragmatic marker and (c) abstract phonological intonation contour.

The output layer represented the human version of the utterance for this sentence which the network had to associate with the input representation. It consisted of an arrangement of the 420 neurons in a graphical matrix of 7 rows by 60 columns. Each column of the matrix represented two time frames of the synthesizer; that is, 20 ms. The seven rows were used to quantize the f0 range of the measured human-produced curves. Figure 6 shows the matrix for one typical sample of each of the eight versions of the sentence.

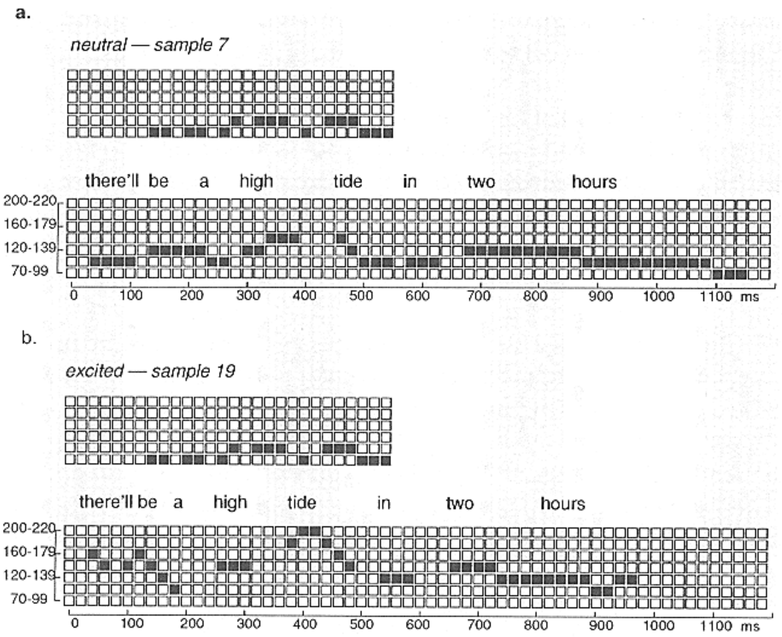

a.
neutral — sample 7
there'll be a high tide in two hours
200-220
160-179
120-139
70-99
0 100 200 300 400 500 600 700 800 900 1000 1100 ms
b.
excited — sample 19
there'll be a high tide in two hours
200-220
160-179
120-139
70-99
0 100 200 300 400 500 600 700 800 900 1000 1100 ms

Figure 5. Two sample input matrices showing the way pragmatic markers, canonical intonation contour and sample number were communicated to the network during the training session. (a) The matrix for the sentence 'There'll be a high tide in two hours' to be paired with sample 7 of the human version uttered neutrally. (b) The same sentence to be paired with sample 19 of the human version uttered with excitement. A dot indicates a non-firing neuron, and *, = and uppercase letters indicate fully firing neurons.

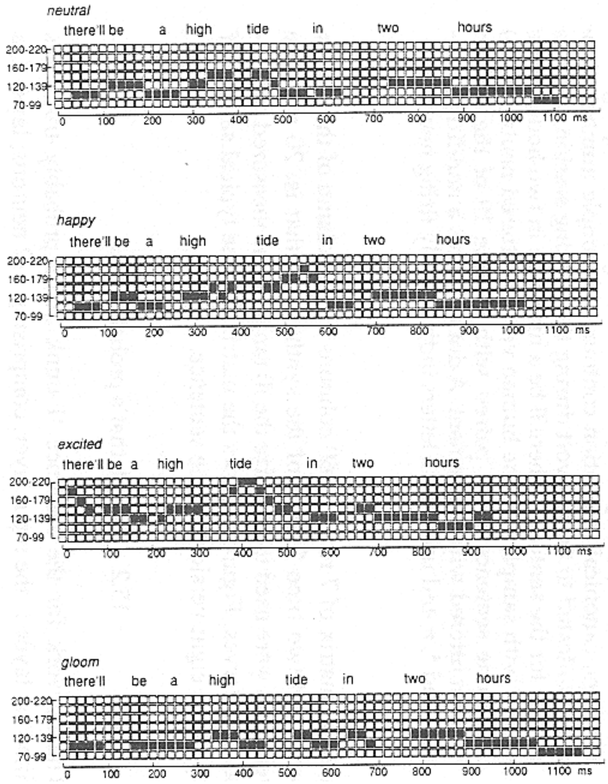

neutral
there'll be a high tide in two hours
happy
there'll be a high tide in two hours
excited
there'll be a high tide in two hours
gloom
there'll be a high tide in two hours
200-220
160-179
120-139
70-99
0 100 200 300 400 500 600 700 800 900 1000 1100 ms

Fig. 6a

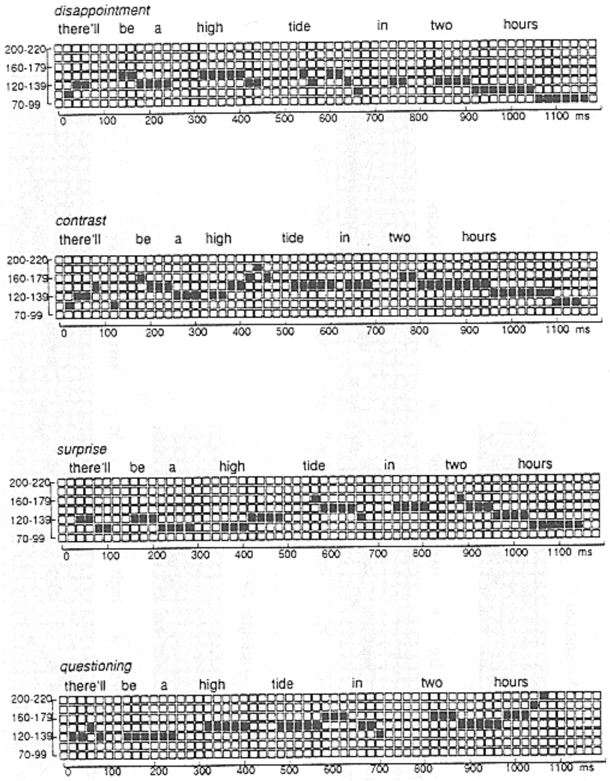


Fig.6b

Figure 6. Matrix representations of the fundamental frequency curves produced by the human speaker for ‘There’ll be a high tide in two hours’. A typical example of each type is shown. Columns in the matrix represent two time frames which were to be used later in re-synthesizing the utterances - each column therefore represents 20 ms. Rows of the matrix represent quantized fundamental frequency in Hz.

15.2 *‘I think that’s probably true.’*

The network for the sentence ‘I think that’s probably true’ was defined as: layer 1, the input layer, comprising 112 neurons; layer 2, a single hidden layer, comprising 75 neurons; layer 3, the output layer, comprising 385 neurons. The program established 8475 connections between the input and hidden layers, and 29,620 connections between the hidden and output layers.

The input layer was set up as a 4 row by 28 column matrix indicating graphically the canonical phonological intonation contour for the sentence, with specific areas showing the pragmatic marker and the sample number. The output layer, representing the fundamental frequency curves produced by the human speaker, was set up as a 7 row by 55 column matrix as before, representing fundamental frequency and time. The matrix had fewer columns than the previous one because the durations of the samples for this sentence were on average shorter in time. Figure 7 shows two sample input matrices.

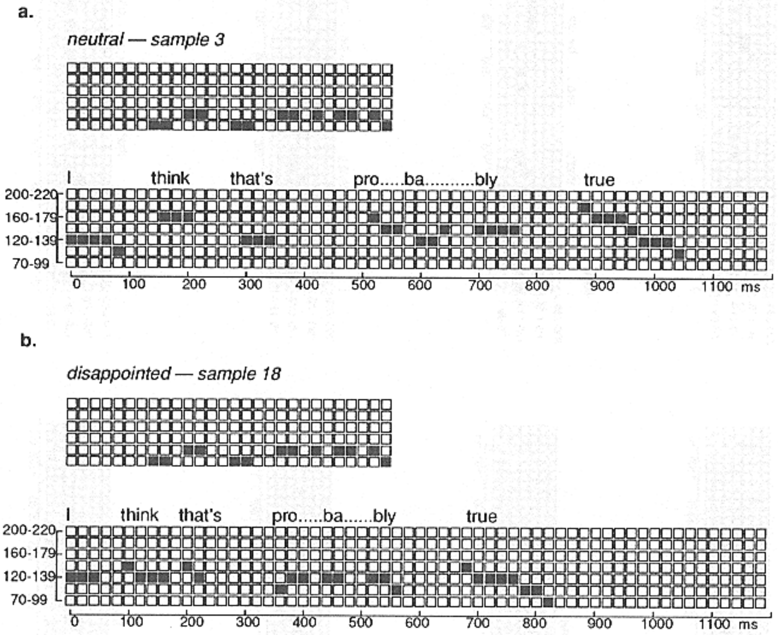

a.
neutral — sample 3
I think that's pro.....ba..........bly true
200-220
160-179
120-139
70-99
0 100 200 300 400 500 600 700 800 900 1000 1100 ms
b.
disappointed — sample 18
I think that's pro.....ba......bly true
200-220
160-179
120-139
70-99
0 100 200 300 400 500 600 700 800 900 1000 1100 ms

Figure 7. Two sample input matrices for the sentence ‘I think that’s probably true’. (a) Pairing with sample 3 of the human version spoken neutrally; (b) pairing with sample 10 of the human version spoken with disappointment.

Figure 8 shows one example of each version of the sentence as represented on the matrix used in the training sessions.

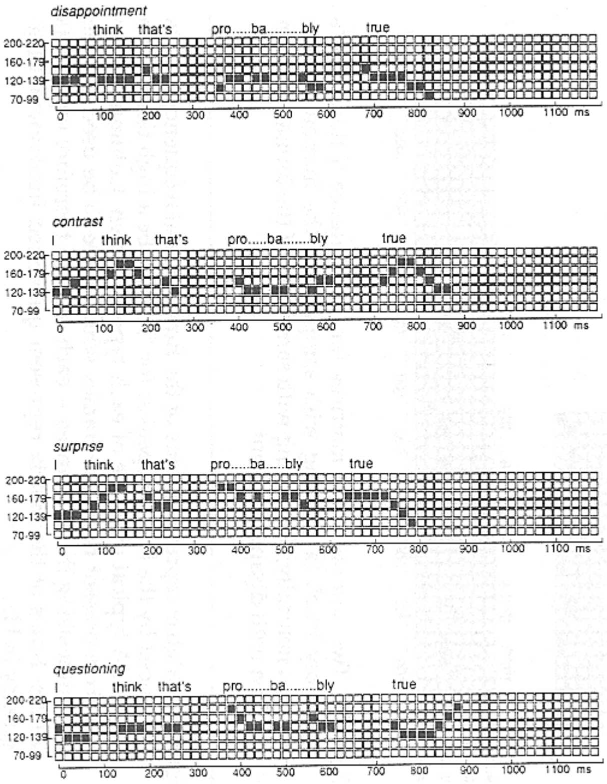

disappointment
I think that's pro.....ba.........bly true
contrast
I think that's pro.....ba.......bly true
surprise
I think that's pro.....ba.....bly true
questioning
I think that's pro.......ba........bly true
200-220
160-179
120-139
70-99
0 100 200 300 400 500 600 700 800 900 1000 1100 ms

Fig.8a

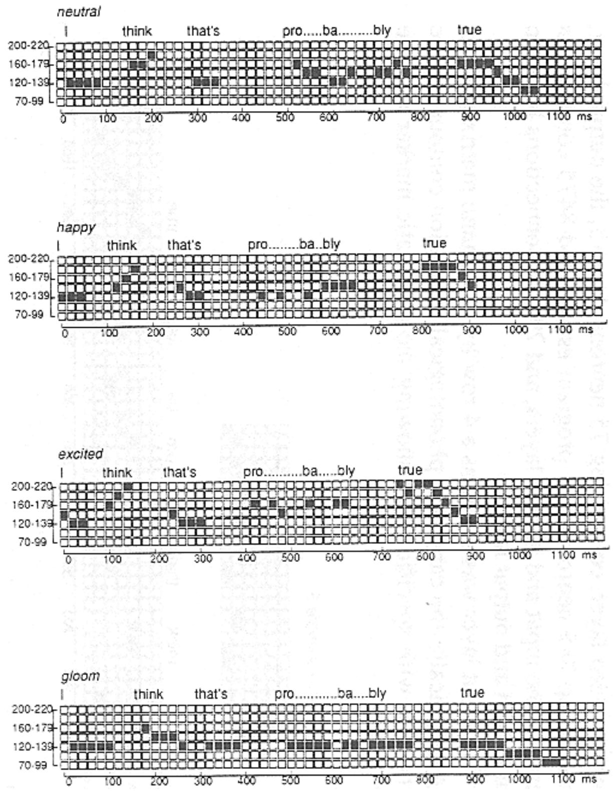


Fig.8b

Figure 8. Matrix representations of the fundamental frequency curves produced by the human speaker for ‘I think that’s probably true’. A typical example of each type is shown. Two synthesized time frames (20 ms) are represented by each column, and rows represent quantized fundamental frequency.

16. THE TRAINING PROGRAM

There were two separate training programs:

a. 16 training sessions in which the networks were trained separately on the 16 different human data sets. This produced 16 trained networks, one for each version of both sentences.

b. 2 training sessions in which the networks were trained on the pooled human data for each of the two sentences, producing 2 trained networks, one for each sentence.

The outcome of program (a) was simply the data reduction referred to earlier. Once the networks were trained, inputs consisting of the phonological pattern together with the pragmatic markers, but without any sample number neurons firing, produced output matrices representing the normalization of the 25 samples in each case.

The networks had no knowledge of the fact that, due to the graphical conventions imposed on the output matrices, maximal firing of neurons was needed to complete the contour pattern and that no more than one neuron should fire in anyone column. Firing of a neuron could be complete or partial; in the system used here firing produced a digit in the output matrix ranging from 0 to 9. Very often more than one neuron was firing in anyone column.

To solve these problems, the output matrices were passed though thresholding rules which counted a neuron as firing only if a value of 5 or more was reached – all other values being reduced to zero (shown by a dot in the figures). At the same time, in cases where more than one neuron was firing in anyone column only the neuron with the highest value was counted, provided that value was 5 or more. Figure 9 shows some examples of the actual output obtained from the networks, together with the results of thresholding.

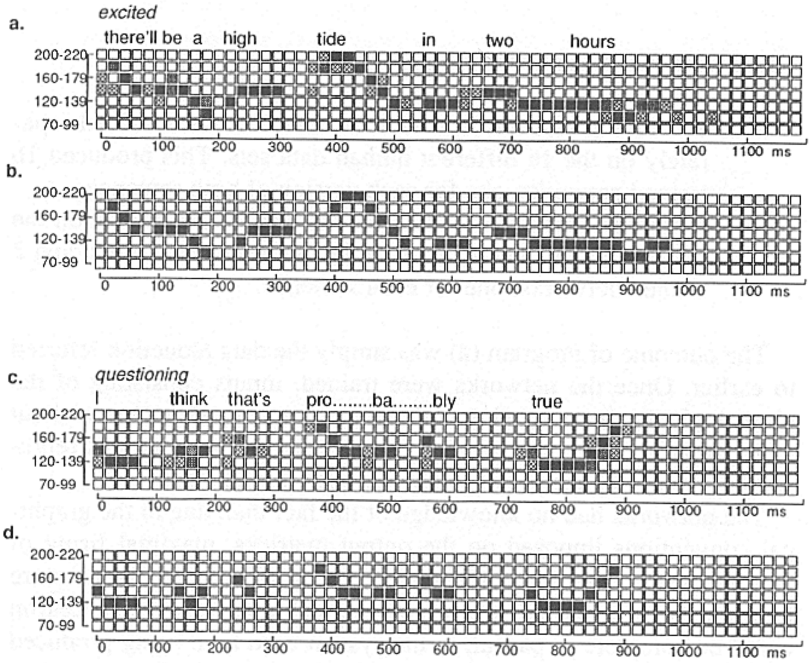


Figure 9. Two actual outputs from the trained networks. (a) and (c) show the outputs before thresholding; (b) and (d) the outputs after thresholding.

Figure 10 shows the final thresholded output of all 16 trained networks when presented with inputs consisting of the canonical phonological intonation pattern together with the appropriate pragmatic markers.

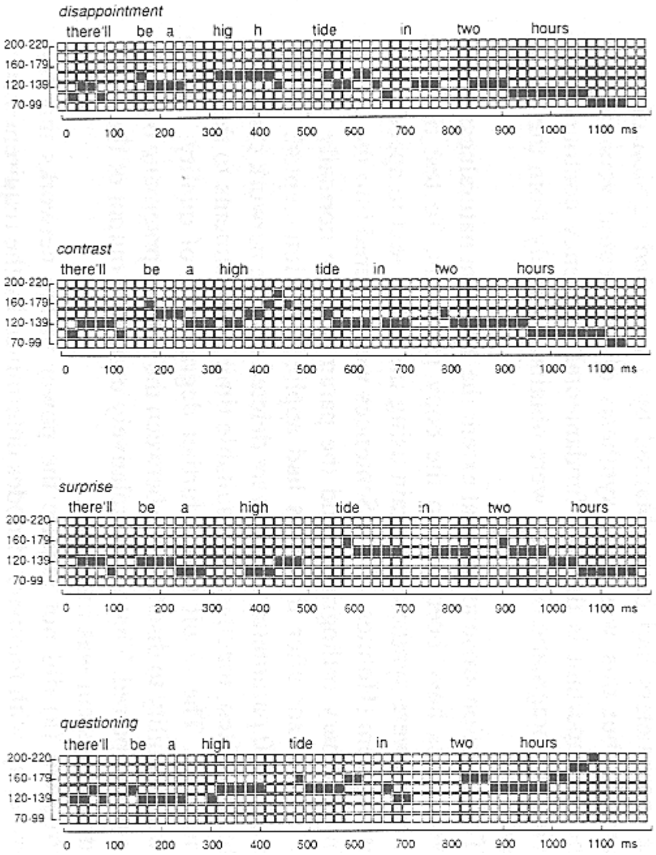


Fig.10a

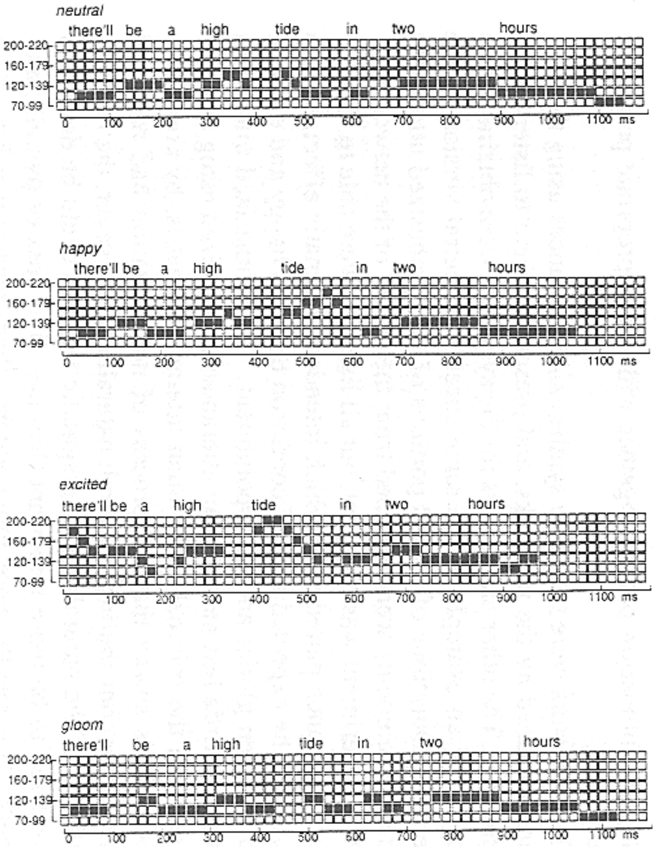


Fig.10b

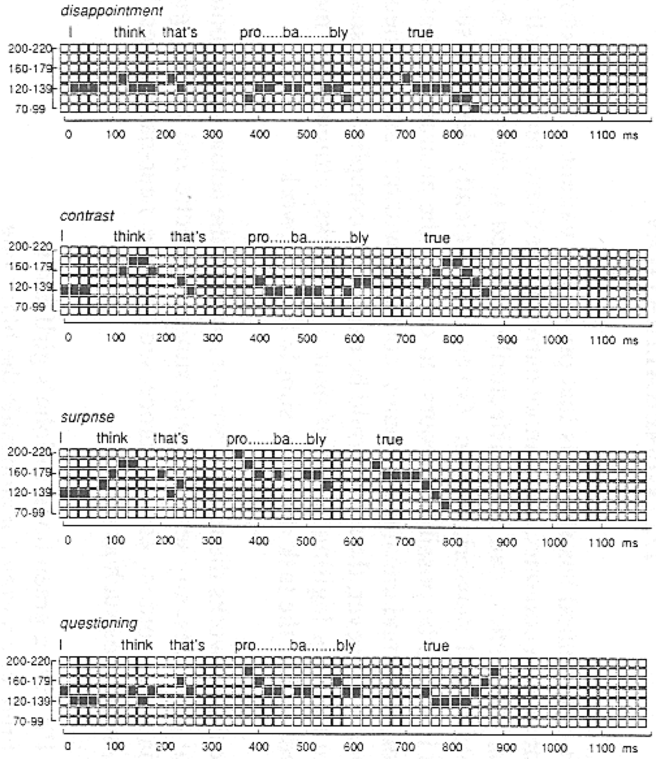


Fig.10c

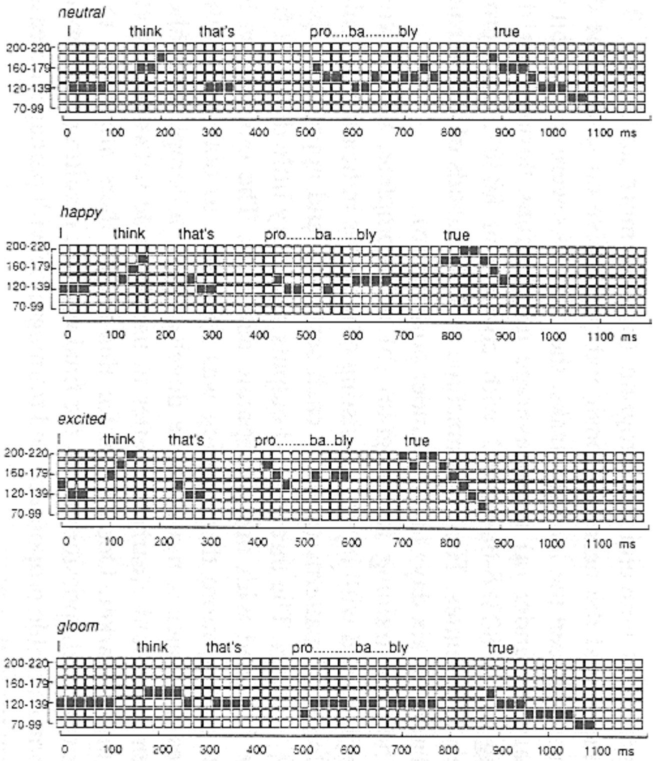


Fig.10d

Figure 10. All outputs from the trained networks after thresholding – eight different utterances for each of 'There'll be a high tide in two hours' and 'I think that's probably true'.

These results were tested by synthesizing utterances using the contours produced by the networks, and presenting them to listeners for evaluation. The other parameters for the synthesis were obtained from analysis of an example of the same speaker's neutral version of the utterance. Durations of voiced portions of the synthesized utterances and the f0 contours were obtained from the output of the networks.

The experiment was informal, but all listeners were able to identify which utterance paired with which intended pragmatic effect, though there was the expected area of confusion between happy and excited, and between gloom and disappointment. The synthesized utterances sounded a little less natural than utterances synthesized using samples taken from the original 400 human utterances (that is, by copy synthesis). This shows that the process of normalization had removed some factor contributing to overall naturalness. However, the fact that synthesis of the naturalness of pragmatic variants could be done was considered to be more important than the slight loss of overall naturalness. The results were more natural than generally achieved in synthesis by rule systems.

Part of the improvement was due to the fact that the other synthesizer parameters had been obtained by a modified version of copy synthesis, but this was copy synthesis of the neutral version only. What is important is that the fundamental frequency contours, conveying the pragmatic overlay, were obtained entirely from the neural networks.

The question arose as to what extent the apparent naturalness might nevertheless have been due to the copy synthesis. To test this, the sentences were synthesized again using the JSRU text to speech synthesis system (Holmes 1988). Sentences were entered into the system using ordinary orthography and the parameter file normally sent to the synthesizer after the rules had applied was intercepted. JSRU-generated f0 parameter values were deleted and the network generated values inserted, together with suitable timing adjustments to the other parameters. The result was a further degradation of quality, but the listeners' ability to discriminate between the various pragmatic overlays was not impaired. That is, the relatively unnatural output of the JSRU synthesis system was improved.

Program (a) did not require the power of the networks that had been set up with numbers of nodes determined by the requirements of the graphical conventions. Program (b) was more ambitious. This program taxed the network considerably. To give an idea of the difference between the learning tasks, the networks were trained on the 25 samples under (a) in an average of 120 runs through the data taking, on a 80386-based PC with co-resident 80387, in approximately 20 minutes. The same machine took thousands of runs through the data over six days to settle under program (b).

When the training for program (b) was complete, the networks were presented with inputs consisting of matrices indicating canonical phonological patterns (one for each sentence) and appropriate pragmatic markers. The thresholded outputs were very little different from those obtained with the 16 separate networks. The output was less clear: that is, more thresholding was necessary to obtain smooth f0 curves. For such a model to be developed for a complete synthesis system, it would clearly be better to use separate networks for each pragmatic marker. One reason for this is that there are going to be more than eight possibilities in the final version, resulting in a network of unmanageable proportions and training time with perhaps increasing ambiguity in its output.

17. GENERALIZATION

What this method has shown is that pragmatic variants of phonological intonation contours, that cannot currently be characterized in theoretical linguistics, can be satisfactorily handled in their detail by a simple neural network system. Given a sufficiently detailed set of rules it would be possible of course to convert the abstract intonation contours to fundamental frequency contours without using a neural network. However, the rules would be extremely complicated. It seems to me this is a legitimate use of neural networks; although, as mentioned earlier, the task was very simple for a network.

Since the networks themselves are not aware of the actual word content of the sentences, any sentence with similar syntactic and phonological intonational content can be processed to give near-natural results. Thus sentences such as:

I'll give a green book to John's girlfriend.

Fred saw a grey cat alongside the black one.

I feel he's usually good natured.

John saw that it's not the best way.

can be dealt with using an interpolation (which can be automatic) across the voiceless gaps in the final f0 contours. The results sometimes show a slight degradation in naturalness, but are still fairly natural when compared with standard speech synthesis systems where the naturalness supplied by emotion is not present.

I believe it would be possible to further generalize by establishing network-derived f0 contours from a relatively small set of canonical phonological intonation patterns. The major problem would be to fit the phonological pattern to the input sentence. The relationship between the sentences and their intonation has been studied in depth by a number of researchers and algorithms have been produced to derive the one from the other (Silverman, 1988). The integration of this work with the work reported here would be the next stage in the application of pragmatic phonetics to speech synthesis systems, either for practical use or as a test bed for perceptual experiments on human speech processing.

18. CONCLUSION

Although neural networks and a speech synthesis system were used to test some of the ideas presented in Part I, the main point of this paper has been the presentation of ideas concerning the role of a selectional subcomponent within a cognitive phonetics. The purpose of the type of subcomponent is to bring together theoretical phonology and pragmatics to provide an input to phonetics. The reason for using synthetic speech was to build a model designed to verify claims made by a pragmatic phonetics that could be linked to traditional phonology. Additionally, it may also be possible to develop a speech synthesis system using neural networks to provide variant pragmatically determined fundamental frequency contours based on the testing experiments reported here.

The core components of theoretical linguistics are not concerned with choices a speaker may make, but with descriptions of his knowledge; nor are they concerned with variants in acoustic fundamental frequency which play a role in communicating to a listener the mood, intentions, beliefs, or feelings of the speaker. Physical phonetics is able to describe these variant fundamental frequency contours but, like theoretical linguistics, does not have the means to choose between them.

Researchers in pragmatics have shown how to handle such phenomena linguistically, but they have not related their descriptions to the physical acoustic signal produced by a speaker. In this paper I have attempted to take a speaker's intention, as signalled by the output of pragmatic descriptions, and use this to overlay the canonical intonation contour characterized by theoretical phonology. The purpose is to integrate these within a cognitive phonetic component to produce appropriately varying fundamental frequency contours in the acoustic signal.

Testing the theory using neural networks in conjunction with a speech synthesis system was designed to illustrate, as general principles, three points:

1. A relationship can be satisfactorily established between static abstract canonical phonological descriptions of sentence intonation together with coded pragmatic information and their associated fundamental frequency contours.
2. The appropriate relationships and generalizations can be trained into a simple neural network to capture what characterizes pragmatic naturalness in human speech, although the detail of the characterization may not be known outside the network itself.
3. A hybrid system of descriptive rules, together with a neural network where the rules are complex or unknown, can be devised whose output can be used to drive a standard synthesizer. The speech output will produce a significant improvement in naturalness that will enable a listener to recognize not just the linguistic meaning of the sentence but also the speaker's intention or feelings.

This test also shows the potential value of neural networks in speech synthesis systems for data reduction with the purpose of adding an element of naturalness to synthesis. It has also shown that a formal method could be devised within a theory of language using the computational techniques of neural networks to relate cognitive or cognitively-oriented descriptions with descriptions of real world phenomena.

NOTE

Synthesis by rule involves producing synthetic speech, using production rules, from a symbolic input such as ordinary orthography or phonetic transcription. Copy synthesis involves an analysis of human speech, either automatically or by hand, and the use of this analysis to directly control the parameters of the synthesizer.

REFERENCES

Chomsky, N. (1965), *Aspects of the Theory of Syntax.* Cambridge Mass: MIT Press.

Holmes, J. (1988). *Speech Synthesis and Recognition.* Wokingham: Van Nostrand Reinhold

Ladefoged, P. (1967, *Linguistic Phonetics, Working Papers in Phonetics* No.6, University of California at Los Angeles.

Levinson, S.C. (1983), *Pragmatics,* Cambridge University Press, Cambridge.

Lyons, J. (1981), *Language and Linguistics.* Cambridge: Cambridge University Press.

Morton, K. (1986), 'Cognitive phonetics – some of the evidence', *In Honor of Ilse Lehiste,* Dordrecht: Foris Publications, pp. 191-194.

Silverman, K. (1988), 'The structure and processing of fundamental frequency contours' Unpublished Ph.D. Dissertation, University of Cambridge.

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

Tatham, M.A.A., (1986) 'The problem of capturing linguistic and phonetic knowledge', *Proc. Inst. Acoustics* 8, 443-450.

Tatham, M.A.A. (1990) 'Cognitive phonetics', In Ainsworth WA. (ed.), *Advances in Speech, Hearing and Language Processing,* Volume I. London: JAI Press, pp. 193-218.