UNDERLYING PHONETIC EXPLANATIONS: REPRESENTING CAUSAL KNOWLEDGE IN A PHONETIC KNOWLEDGE BASE*

A. J. H. Simons

Department of Computer Science,
University of Sheffield, SHEFFIELD S10 2TN.

ABSTRACT

It is suggested in this position paper that phonetic knowledge can be put to much more effective use in automatic speech recognition than has been achieved to date. Reported phonetic expert systems typically use production-rules; hypotheses about broad classes of sounds are progressively refined; the systems reason with tokens such as acoustic cues and linguistic features. Following an established methodology for the design process of expert systems, the scope of the problem is redefined to show the types of phonetic knowledge which a true expert will have to model. The conceptualisation of phonetic knowledge is reappraised with a view to providing appropriate symbolic tokens, defined in markedly different domains, for reasoning in causal explanations rather than surface classifications. The kinds of formalisation required are discussed, showing the degree of detail necessary and the interactions within and across the different knowledge planes.

INTRODUCTION

Waterman [1], citing Buchanan et al [2] identifies several stages in the design and implementation of expert systems. The first three of these are: identification, conceptualisation and formalisation. During the identification stage, the scope of the problem is often narrowed down to one sub-area where the ideas for a demonstration prototype can evolve and be tested. During the conceptualisation stage, key concepts, relations and control mechanisms are found and the issue of 'grain size' is addressed. During the formalisation stage, a knowledge representation framework is chosen. A major problem in expert system building is that, as a result of incremental development, a system prototype can become unwieldy, leading to its abandonment in favour of a substantially different model (a complete 'paradigm-shift'). The results of going through the above stages, re-addressing the fundamental issues of knowledge representation, can then look quite different.

It is reasonable to expect that expert systems for automatic speech recognition will not be immune from this difficulty. Current system prototypes (see, for example, papers in section 23, Proc. I.C.A.S.S.P., 1986) would seem to have a limited lifespan, given this experience in other fields of expertise. The author suggests that the scope of the recognition problem will have to be expanded to model many more of the sources of dependency-governed variability. Available tokens or concepts do not seem to suffice. Not only do we need quantifiable acoustic measurements, but we need to map these onto a qualitative vocabulary in terms of acoustics, descriptions of articulator movements, syllable structure, speaker type and possibly further areas of linked knowledge. The formalisation of such a speech understanding system will probably differ from standard rule-based architectures. An object-oriented model [3] suggests itself, in which the various domains of phonetic knowledge are represented by networks of structured objects.

*Paper presented at the Institute of Acoustics Conference on Speech and Hearing, Windermere, 30 November 1986. Proc. Inst. of Acoustics, Vol. 8, No. 7, 499-505, (1986).

IDENTIFICATION: REAPPRAISING PHONETIC KNOWLEDGE

In Buchanan's methodology, 'identification' means isolating a specific problem-area well suited to expert problem-solving approaches. In automatic speech recognition (ASR) and more specifically in acoustic-phonetic decoding this has indeed been attempted. In Huckvale's terms [4] phonetic variability is partly controlled (by limiting the speaker, rate of speech and environment conditions), partly accommodated (by choosing robust threshold-governed algorithms to group data into broad phonetic classes), and modelled only in that the different sets of preconditions for rules express the 'cases' of some other external determining factor.

The field of expertise which current systems actually tackle is phonetic context-dependency. This has most often been conceptualised as a set of allophonic and phonotactic rules [5], which exploit the information-bearing features of allophones to constrain the identities of the segment under consideration and its immediate context. An approach based on feature extraction has therefore emerged: see, for example, Zue et al [6] and reports from the Edinburgh EUSIP team in this volume - SEGLAB. At the lowest level acoustic cues are sought, as evidence for the existence of higher-level categories, identified by their phonetic or linguistic labels. The terminology varies between systems, but generally there is a division between 'acoustic features' and 'phonetic features', corresponding to the two levels of analysis.

The ability to use this kind of phonetic knowledge in this way is to some extent a consequence of the systems' self-imposed limitations relating to the kinds of variability expected; this is often acknowledged. 'Real' speech recognition will require the modelling, in Huckvale's sense, of many more causes of variability. Worden et al [7] note that in many task domains, similarly limited expert systems are being offered as 'expert assistants', since they lack the breadth of knowledge, the available data and the inference methods to perform their task autonomously. For most speech applications such an 'assistant' would be inappropriate. If the task is to be attempted at all, the whole task must be attempted.

In view of this, 'identification' of the problem for a phonetic expert system must be taken to mean recognising all the causes of variability in speech, together with the knowledge that a phonetician uses when interpreting speech data. Huckvale [op. cit.] describes the former; below I attempt to outline the sources of knowledge available to a phonetician. These, some as yet unexploited, range from those which reside firmly in empirical domains to those of a more abstract nature.

PHONETIC KNOWLEDGE SOURCES

- Categorical - knowledge of the minimum set of contrastive sound categories (phonemes) of a language; the ability to associate any sound with one or other category and to modify expectations based on already instantiated categories.
- Linguistic-Phonetic - explicit knowledge of the fine phonetic quality of allophonic variants of phonemes: intrinsic allophones determined by immediate context (admitting of articulatory or acoustic explanation), extrinsic allophones linguistically distributed and archiphonemes in positions of neutralisation of contrast; strategies for mapping from some representation of phonological oppositions onto expectations of fine phonetic qualities (predicting allophonic variation).
- Articulatory - knowledge of the processes of speech production, the likely positions and movements of various articulators and how this affects retentive and anticipatory coarticulation; absolute and dynamic physical constraints including the shape and size of cavities and effects of speaking-rate; knowledge of the mean and range of articulatory configurations; the abililty to relate the place and manner of articulation to acoustic events (causal knowledge) including possible articulatory compensations admitting of acoustic explanation.
- Acoustic - knowledge of the acoustic characterisations of speech sounds, including mean correlates of all major variant (allophonic) classes; absolute acoustic constraints in time/frequency domains; the masking and interfering effects of different types of acoustic energy (causal knowledge).

- Visual - knowledge of the types of visual objects present on spectrograms and their acoustic/articulatory correlates; a sophisticated ability to perceive patterns and identify the behaviour of visual objects acting conjointly.
- Syllabic - knowledge of the dependencies existing between syllable structure and the articulatory/acoustic realisation of segments (causal knowledge).
- Prosodic - knowledge of the stress and intonation patterns of language; the dependencies existing between this and other domains (causal knowledge).
- Speaker-specific - ability to adapt expectations based on some internal model of speaker type and voice quality.
- Statistical - knowledge of the a priori and context-dependent probabilities of various segments or clusters.
- Parametrical - robustness of evidence obtained from various signal-processing techniques: absolute and relative constraints of the hardware and representation used.

A proper exploitation of these available knowledge-rich domains should enable the tackling of phonetic variability in all its forms. If the problem is viewed in the above terms, it can be seen to exist on many levels; each level has its own rules governing well-formed behaviour. The key notion is one of causality: changes at one level may have a non-linear effect on other levels; these effects may be propogated further from level to level.

CONCEPTUALISATION: TOKENS FOR REASONING

Approaches based on feature-extraction reason on the basis of the presence, absence or combinations of these features. For example, Johnson et al [8] have the following Prolog rules for identifying fricatives on spectrograms:

```
event(fricative,T1):- type_of_pattern(fuzzy,T1),
        not(length('<9',T1)).
event(strident,T1):- event(fricative,T1),
        intensity(high,T1).
event(alveolar,T1):- event(strident,T1),
        cut_off('2700',T1).
```

meaning that, if at a given point in time there is a fuzzy pattern not shorter than 9 units, it is a fricative; if the fricative is high in intensity it is strident; if the visible energy of a strident event disappears below 2700 Hz, it is alveolar. Key concepts are therefore the feature (visual, acoustic and articulatory descriptions), thresholded numerical information and the heuristically-expressed phonetic deduction rule.

Such a conceptualisation tries eventually to map all phonetic variability onto a large set of phonotactic rules. This has a profound influence on the way the knowledge-base must be structured and interpreted; and on what kinds of phonetic reasoning may be accomplished. Zue groups his acoustic-phonetic knowledge into necessary, sufficient or redundant cues [9]; this must result in a strategy that fires rules in order of their 'necessity'. Sometimes extra preconditions are included to alter this behaviour [3] in the light of other contextual knowledge. EUSIP (reported in a personal communication) find that a suite of algorithms for detecting one feature generally returns multiple results where the most stringent algorithm reports only occasionally, but with a high confidence factor, and the more lenient algorithms report often, but frequently make misidentifications. Glass and Zue [10] note similar 'impostors', or misidentified features, in a nasal-detector: the strategy becomes one of refinement, or another layer of rules to limit the likely set of 'real' features among the impostors. The status of the feature is therefore considerably weakened; correspondingly the rule-base becomes more difficult to control and interpret.

It is not so much that phoneticians carry around in their heads an exhaustive set of rules for mapping every phonotactic case onto acoustic cues; rather that they understand the underlying causal processes which bring about the highly dependent surface variations. A phonetician can reason in terms of the movements and

configuration of articulators; or talk of a 'fronted [k]' without reference to parametrical measurements and follow this through to its acoustic consequences. He can expect to find well-represented information in stressed syllables; he can predict coarticulation. Here, then, important concepts are causality and qualitative descriptions.

Qualitative reasoning has been exploited to simplify flow-control, circuit diagnostics and to model naive physical systems [11]. Clocksin and Morgan [12] note that although quantitative techniques (in fluid control), if 'tuned', always provide more accurate results, qualitative models are more robust, data-independent and react appropriately in emergencies(!) Causal models are being exploited in the field of automatic learning - Van de Velde [13] notes that 'deep knowledge' expressed in a causal way is more accessible than surface heuristics. A causal explanation in phonetics, then, is of the form: 'this acoustic realisation was caused by that particular articulatory configuration executed with this particular prominence by that particular speaker'. For a machine to be able to reason in such a qualitative and causal way, it will require symbolic tokens (concepts) with which to build a description of standard behaviour (relations) in each domain understood by the system.

The inadequacy of (linguistic) distinctive features for this purpose is discussed in Simons, 'Phonetic tokens for symbolic reasoning', in this volume. Fine acoustic measurements may be made, but these are usually expressed in terms of energy ratios, frequency values and so on. Also, they are prone to error where they attempt an early interpretation of data, for example a 'nearest neighbour' algorithm deciding the identity of formants from an LPC scatter-plot (EUSIP data). It would be better to have a small qualitative vocabulary, applicable to many speakers, since such spectral measurements are often too specific, or knowledge-poor. Conversely, articulatory descriptions can be given in general terms, but these are too broad. Here, what is needed is, as it were, a 'quantum representation' of articulator movements. Similar descriptive frameworks could be devised for other domains.

FORMALISATION: ACTIVATING KNOWLEDGE SOURCES

Production rules have often been chosen for reasons of programming convenience and the apparent ease with which human judgemental expertise can be rapidly encoded. Prolog is used directly [8] or indirectly [14]; a Mycin-based [15] expert system shell may be used [6]; or rules written directly as Lisp functions (EUSIP). While this approach enables rapid prototyping, the disadvantages of production rules (see Jackson [16] for a full comparison) would seem to outweigh the advantages when it comes to implementing full-scale systems. Briefly, the prior assumptions behind the archetypal rule-based system do not necessarily hold true for (proposed) large speech recognition systems. The functional independence of rules cannot be guaranteed (mutually dependent sources of variability must affect each other); the control regime will most likely need to re-specify its own metarules (changing strategy to follow through the consequences of reversing dependencies); the unique context of application criterion will generate hundreds of almost similar rules competing for the attention of the rule interpreter (attempting to isolate each 'case' by adding preconditions to rules).

Allerhand and Fallside [17], [18] have implemented a hybrid recogniser which makes use of a syllable grammar to express declaratively knowledge about stressed and unstressed environments. The success of this is that the grammar encodes in the one compact, context-free, knowledge-source information which, had it been expressed as allophonic and phonotactic rules, would have appeared in a disjunctive form as multiple cases for different phonetic categories. Their intuition is that a syllable model provides the better, more general form of causal explanation.

This principle can be further extended in an object-oriented model. Phonetic knowledge is expressible in several different domains. Each domain has its own associated causal explanations for speech events. Explanations of the speech process may be seen to lie mainly in one or other domain for consecutive time segments depending on the nature of the speech event. An expert system should understand which domain provides the best explanation at any given point, and should be driven by that domain (the 'active domain'). Where a domain fails

to provide any explanation, mappings from one domain to another should be considered and control should pass to the most 'active' knowledge source.

As an example of this, consider the articulatory domain. Instead of labels such as 'velar', 'alveolar', it would contain structured objects exemplifying frame-like [19] descriptions of articulatory configurations.

```
NAME:   artdesc.vowel./a/        (* articulatory /a/)

SUPERS: artdesc.vowel            (* inherits from vowel)

COMPS:  NIL                      (* has no components)

CAUSE:  acoudesc.vowel./a/       (* causes acoustic /a/)

ALV:    1        (* alveolar influence)
VEL:    0        (* velar influence)
PHA:    2        (* pharyngeal influence)
AP:     3        (* jaw aperture)
```

FIGURE 1: SIMPLIFIED FRAME FOR AN ARTICULATORY DESCRIPTION

These frames would use class inheritance for commmon information. They would have sequence links showing allowable progressions from one configuration to another, and attached procedures to simulate coarticulation frames. They would have causal links connecting with acoustic frames. Where the articulatory event is essentially dynamic, or a sequence, this would be represented by pointers to sub-frames with the appropriate sequence links specified between these.

In a fully-developed model, acoustic data would instantiate progressively higher-level acoustic frames until the data-driven phase exhausted itself; during this mappings would be made to other domains, which would supply consistency-checking information. The most consistent domain model would then initiate a model-driven phase looking for patterns to match against data, including other consistent domain modelled data.

CONCLUSION

There would seem to be great benefit in exploiting the full potential of second-generation expert systems in ASR. The keynotes are the ability to reason causally and qualitatively. The advantages of an object-oriented system are the facility to express knowledge declaratively and the facility to model domain behaviour explicitly.

ACKNOWLEDGEMENT

This work was supported by ALVEY grant MMI 052.

REFERENCES

[1] D. A. Waterman, 'A Guide to Expert Systems', Addison Wesley, 135-141, (1986).

[2] B. Buchanan et al, 'Constructing an expert system' in 'Building Expert Systems', F. Hayes-Roth, D. A. Waterman and D. Lenat (eds.), Addison Wesley, (1983).

[3] A. Goldberg and D. Robson, 'Smalltalk-80: the language and its implementation', Addison Wesley, (1983).

[4] M. A. Huckvale, 'Modelling acoustic and phonetic variability of speech', Proc. I.E.E. Int. Conf. on Speech Input/Output; Techniques and Applications, 54-58, (1986).

[5] K. W. Church, 'Allophonic and phonotactic constraints are useful', Proc. I.J.C.A.I., 636-658, (1983).

[6] V. W. Zue and L. F. Lamel, 'An expert spectrogram reader: a knowledge-based approach to speech recognition', Proc. I.C.A.S.S.P., Paper 23.2, (1986).

[7] R. P. Worden, M. H. Foote, J. A. Knight and S. K. Andersen, 'Co-operative expert systems', Proc. E.C.A.I., 319-334, (1986).

[8] S. R. Johnson, J. H. Connolly and E. A. Edmonds, 'Spectrogram analysis: a knowledge-based approach to automatic speech recognition', in 'Research and Development in Expert Systems', ed. M. A. Bramer, British Computer Society Workshop Series, C. U. P., 95-103, (1985).

[9] M. A. Bush, G. E. Kopec and V. W. Zue, 'Selecting acoustic features for stop consonant identification', Proc. I.C.A.S.S.P., Paper 16.8, (1983).

[10] J. R. Glass and V. W. Zue, 'Detection and recognition of nasal consonants in American English', Proc. I.C.A.S.S.P., Paper 51.5, (1986).

[11] G. Cohn and P. Hayes, 'Qualitative reasoning', tutorial given on 21 July at E.C.A.I., Brighton, (1986).

[12] W. F. Clocksin and A. J. Morgan, 'Qualitative control', Proc. E.C.A.I., 350-356, (1986).

[13] W. Van de Velde, 'Explainable knowledge production', Proc. E.C.A.I., 8-21, (1986).

[14] P.-E. Stern, M. Eskenazi and D. Memmi, 'An expert system for speech spectrogram reading', Proc. I.C.A.S.S.P., Paper 23.1, (1986).

[15] E. H. Shortliffe, 'Computer Based Medical Consultations: MYCIN', American Elsevier Publishing Co., (1976).

[16] P. Jackson, 'Introduction to Expert Systems', Addison Wesley, 29-92, 217-220, (1986).

[17] M. H. Allerhand and F. Fallside, 'A hybrid recogniser for speech patterns', Proc. I.C.A.S.S.P., Paper 23.6, (1986).

[18] M. H. Allerhand, 'A knowledge-based approach to phonetic decoding in continuous speech', PhD Thesis, Cambridge, (1986).

[19] M. Minsky, 'A framework for representing knowledge', in 'The Psychology of computer Vision', ed. P. H. Winston, McGraw-Hill, (1975).