PHONETIC TOKENS FOR SYMBOLIC REASONING: AN APPRAISAL OF DISTINCTIVE FEATURE THEORIES*

A. J. H. Simons

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

ABSTRACT

Phonetic expert systems for automatic speech recognition need to be able to reason symbolically with tokens corresponding to natural classes in the acoustic, articulatory, auditory or linguistic domain. Distinctive Features have traditionally fulfilled such a function in linguistics. This paper describes the roles that symbolic reasoning tokens would be expected to play in an expert system and examines the pedigree of various descriptive and feature systems in this light. Traditional phonetic descriptions and three well-known Distinctive Feature systems are discussed. They are found wanting in various respects: most significantly, they fail to provide a fine-grained representation amenable to a causal explanation of underlying phonetic behaviour. Preliminary examples of symbolic reasoning tokens appropriate to the articulatory domain are proposed, together with a schema for integrating them in a knowledge representation.

INTRODUCTION

In the design of phonetic expert systems for automatic speech recognition (ASR), various roles have been proposed for symbolic tokens corresponding to features borrowed from linguistics. In recent American, British and French systems, features taken from the acoustic, phonetic, linguistic and perceptual domains have been advanced. Their purpose has been, broadly speaking, twofold: to characterise gross phonetic classes and thereby to aid the process of recognition by the progressive refinement of these classes [1] (and see also reports in this volume from the Edinburgh University Speech Input Projects); and to evaluate the performance of recognisers in terms of a feature-count for misidentifications [2], [3]. A third strand is emerging which, in the opinion of the author, will surpass the other uses to which features are put, namely the devising of tokens amenable to symbolic reasoning processes and capable of providing transparent explanations of these.

The bench-mark for any system capable of detailed phonetic reasoning is the performance of a trained phonetician. The strategies employed by a phonetic expert during spectrogram reading experiments have been investigated by protocol analysis [4], [5]. One interpretation of similar expert behaviour (in different domains) held by Waterman [6] would seem to apply also to a phonetician. In the trivial case, he will express his knowledge in the form of a surface heuristic, whereas given a difficult or novel problem he will resort to first principles, reasoning in qualitative terms using a causal model of the highly interdependent acoustic, articulatory and prosodic domains.

Clearly, it would be invaluable to obtain a descriptive system that could cope with reasoning at this degree of granularity. An analysis of existing descriptive and feature systems given below illustrates their deficiencies in this respect. Later, some suggestions are made towards developing an appropriate set of symbolic reasoning tokens, giving examples.

*Paper presented at the Institute of Acoustics Conference on Speech and Hearing, Windermere, 30 November 1986. Proc. Inst. of Acoustics, Vol 8, No 7, 507-513, (1986).

Classification in terms of place and manner of articulation for consonants and height, depth and rounding for vowels has long been the traditional method of phonetic description, exemplified in the IPA consonant chart and the cardinal vowel quadrilateral. The knowledge explicit in such a description is, as it were, a recipe for a phonetician to produce a given speech-sound; it is expressed in terms of the positions and movements of articulators to achieve certain phonetic targets. While widely useful in descriptive linguistics, it is unreasonable to expect automatic systems to be able to handle, from a machine's viewpoint, such an economical and actually complex phonetic notation as it stands. This is because the ability to describe in traditional phonetic terms presupposes an ability to interpret, reason and discriminate at a much finer level; these processes are neither wholly explicit nor represented systematically in the traditional system.

The strengths of the traditional system are twofold: it provides a rich qualitative descriptive framework; and it is impressionistic at heart. A phonetician can put the variable qualities of a velar burst down to coordinated articulatory gestures, and talk in terms of a "fronted [k]" without reference to absolute measurements or exhaustive sets of phonotactic rules. He may specify any degree of phonetic detail 'on the fly', governed by other local or higher-order requirements. To model this kind of behaviour would require causal knowledge and qualitative reasoning techniques.

Knowledge-based ASR machines that operate on a feature count are making, in general, simple quantitative decisions (the presence, versus the absence of a linguistic feature). To reason qualitatively with traditional phonetic descriptions presupposes a systematic quantification of phonetic knowledge down to the finest degree of granularity. The traditional categories would need to be, as it were, "decompiled". Furthermore, machines would need sufficient detail in their knowledge representations to be able to derive automatically the appropriate selectivity corresponding to a phonetician's impressionistic behaviour.

The traditional system as it stands is not without faults. Vowels and consonants are described using different paradigms - from a machine's viewpoint it would be desirable for these to be unified, especially when giving consistent accounts of coarticulation. Conflicts with articulatory evidence have been discovered since the establishment of the traditional model. For example, the open/close and front/back distinctions are not as categorical as the cardinal vowel quadrilateral might lead one to suppose. Ladefoged [7 p.70], [8 p.66], finds that the notion of tongue height is problematical and many traditional separations along the open/close axis are often neutralised or even reversed in one case. Rossi [9] finds classification along a front/back axis counter-intuitive and misleading in explanations of real muscular activity.

DISTINCTIVE FEATURE THEORIES

Developed originally by observing the non-uniqueness of phonemic solutions to problems in phonology, distinctive features provide a systematic framework for modelling phonological processes. Their proven usefulness in phonology is not the issue in this discussion (the various arguments for and against individual theories are well known in linguistics); rather their suitability for expressing the kinds of detailed phonetic knowledge required by ASR machines.

The feature theories discussed are those of Jakobson, Fant and Halle (JFH) [10], Chomsky and Halle (CH) [11] and Ladefoged (LAD) [7], [12]. Of these, JFH is an acoustic classification based ostensibly on the distribution of energy in the spectrogram, CH is based largely on schematised articulatory distinctions; both provide a binary, non-redundant framework. LAD admits traditional phonetic descriptions in both acoustic and articulatory domains as part of a multi-valued system with a certain amount of redundancy.

Properly construed, distinctive features are abstract tokens used in the demonstration of phonological processes, rather than fine-grained phonetic tokens for the explanation of real physiological and aerodynamic processes. They are designated 'features' for their sub-phonemic granularity and 'distinctive' because their main purpose is to distinguish phonemes minimally by the presence or absence of a single feature. This more or less accounts for

their attractions and disadvantages at the outset: they appear to be computationally efficient, they may represent perceptually salient categories; but the classificatory philosophy behind their definitions undermines their explanatory power.

From the computational viewpoint they offer an apparently easy data structure (a matrix of features, usually binary). They unify the representation of consonants and vowels. They attempt to capture all the significant linguistic generalisations within (JFH) and across (CH, LAD) languages.

Singh finds that CH may have some correspondence with complex perceptual categories: Singh [13] and Singh & Woods [14] use two multi-dimensional scaling techniques, INDSCAL and MDSCAL, to divide the space of phonemes according to their natural separability (this derived from confusion matrices). These statistical processes divide repeatedly a group of objects, thereby defining an axis or dimension along which the subgroups are opposed. The interpretation of Singh's dimensions as CH oppositions, and therefore amenable to Chomskyan explanation, should not be assumed lightly, however: the technique provides no proof of this.

The features' discriminatory function leads increasingly to less tractable choices of boundaries between contrasting or opposing pairs, as the phoneme-space is divided more and more finely. So whereas the first few divisions may be seen to rest on robust articulatory or acoustic evidence, subsequent divisions rely on actually more complex linguistic notions (tense, lax) and even contentious theoretical decisions (classification, binarity, one-to-one mapping), giving rise to controversy over the supposed naturalness, from our point of view, of the phonetic classification thereby obtained. At worst, distinctive features correspond to ad hoc threshold boundaries: any suitable feature set which achieved the appropriate phonemic contrasts might be chosen.

More importantly, there is no guarantee that features so obtained can serve as tokens of explanation for the underlying real acoustic or real articulatory behaviour. CH found that phonological rules operated with maximum efficiency on three degrees of vowel-height (contrasting with the IPA's four) by making use of the tense/lax dimension to divide the vowel-space further. Apart from the fact that Ladefoged [12 p.262], finds it intuitively preferable to posit four degrees of vowel height for Danish and maybe only two for Arabic, instrumental evidence (cited above) tends to put these classic notions of vowel height into doubt.

The appropriateness of feature definitions to their stated domain is, in certain cases, doubtful. So, while the original JFH definitions of compact/diffuse and acute/grave are well motivated from the acoustic standpoint, other definitions may look more like attempts to capture articulatory notions such as [+sharp] for palatalisation, [+flat] for velarisation, labialisation or pharyngealisation. Similarly the deliberately economical CH division of the vocal tract using features anterior, coronal, high, back and low is derived most obviously from articulatory parameters. However other features may seem to be begging definition in other domains like [+sonorant] which rests on a belief in a unique setting of the vocal cords in which spontaneous voicing is possible (Ladefoged prefers an acoustic definition [12 p.261]), or the tense/lax distinction which rests on a belief in heightened muscular effort for tense sounds. The tense/lax opposition is almost certainly a complex linguistic distinction, not directly predictable from articulatory or acoustic observations.

If binary and non-redundant, a feature-system creates unnatural classes (there is an inevitable trade-off in binary systems between economy and naturalness). For example, CH would have /p/ and /k/ related by the shared feature of [-coronal]. This gives rise to the counter-intuitive explanation "they are similar because they are not articulated with the tip of the tongue". More seriously, binarity leads to the formal similarity of rule progressions, of which one may be properly ordered and the other not [7 p.103].

If multivalued and redundant, the admission of features of different orders tends to unbalance the system in terms of the weight of, or significance to be given to, any particular feature. LAD maps feature values onto a percentile scale (admitting a good deal of guesswork) for comparison. This means that one feature, like nasality, may be present (100%) or absent (0%) entirely and another feature may be present to a degree, like place, where

the alveolar place of articulation is rendered by [place 85%] ie 85% of the normalised distance from glottis to the lips. Thus the large variation in percentile values of true binary oppositions will tend to outweigh the more subtle multivalued features.

To a certain extent, there is an implicit hierarchy of features in any system and none of the three systems represents this explicitly, although there is a pointer in this direction in Ladefoged's discussion of 'cover features' [7 p.109].

DESIGN REQUIREMENTS FOR SYMBOLIC REASONING TOKENS

By using the term 'token' I wish to break away from the notion of distinctive features. Tokens should function as units of knowledge representation, rather than of classification, and should reside firmly in observable domain behaviour. Symbolising the finest-grained knowledge available to machine reasoning processes, it is envisaged that each token will have a small range of values corresponding to the degree of the token's influence in the higher-level unit, or descriptor [15].

Tokens should be defined in terms of the domain in which their explanatory value is most transparent. Descriptors should be organised hierarchically, weighted and integrated with the reasoning-mechanism for that domain. Mapping between domains should be non-linear: a corollary of this is the possibility of multiple, or non-unique phoneme classifications in a given domain.

Descriptors should provide for at least these three functions: simulation, inferencing and explanation. Simulation is where, for example, allophonic variation can be modelled in explicit detail showing the effects of small adjustments to the behaviour of articulators; inferencing is the process whereby contiguous states within and across domains can be predicted using various constraints including causal knowledge; explanations should be given of the process of identification based on a model of a phonetician's causal knowledge and these should be amenable to a balanced evaluation metric for system performance.

The implementation of static and dynamic descriptors would make use of a frame-like data structure [15], where the 'slots' contained instantaneous information (the values of tokens), contextual information (sub-frames) and dynamic information: relationships between sub-frames expressed qualitatively as the values of tokens.

LIMITED DOMAIN EXAMPLE

Following research into tongue and jaw activity (Neary [16]), a polar coordinate description of tongue position (Coker & Fujimura [17], Mermelstein [18]), coordinated action of tongue muscles (Zerling [19]), Rossi proposes three dissymetrical influences dominated by the action of certain muscles for a systematic articulatory description of vowels [6]:

pharyngeal influence: hyoglossus activity: {jaw opening}
velar influence: styloglossus activity: {lip protrusion}
alveolar influence: genioglossus activity: {spread lips}

Adapting his table of seven binary features for cardinal vowels to allow a small range of qualities with fewer tokens, and applied to RP instead, this gives one possible static descriptor representation:

TABLE 1.

vowel(MRPA)	alv	vel	phar	aperture
ii	4	0	0	1
i	2	0	0	1
e	2	0	1	2
a	1	0	2	3
aa	0	0	4	4
o	0	1	4	3
oo	0	2	3	2
u	0	2	0	1
uu	0	4	0	1

This table is a first estimation; further changes in the number and ranges of tokens will certainly be necessary.

A major improvement of this kind of representation is that it rests in real, observable muscular activity. Compensatory effects could be modelled by multiple articulatory representations of the same acoustic-phonetic event. Cross-domain influences may not be linear, but here, for example, there is a strong correspondence between the pharyngeal influence and the peaking of F1 and F2 in the lower portion of the spectrogram (the 'single equivalent formant' effect).

CONCLUSIONS

Although distinctive features would appear to supply appropriate systematic information for ASR systems, this is not properly the case. Much further investigation is needed to provide symbolic tokens truly representative of behaviour in particular domains of speech.

ACKNOWLEDGEMENT

This work was supported by ALVEY grant MMI 052.

REFERENCES

[1] D. P. Huttenlocher and V. W. Zue, 'A model of lexical access from partial phonetic information', Proc. I.C.A.S.S.P., paper 26.4, (1984).

[2] J. H. Connolly, 'A multi-level functional-linguistic approach to the evaluation of speech recognition systems', Proc. I.E.E. Int. Conf. on Speech Input/Output; Techniques and Applications, 129-133, (1986).

[3] P. J. Roach, H. N. Roach and A. M. Dew, 'Assessing accuracy in automatic identification of phonetic segments', Proc. I.E.E. Int. Conf. on Speech Input/Output; Techniques and Applications, 216-219, (1986).

[4] V. W. Zue, 'Acoustic-phonetic knowledge representation: implications from spectrogram reading experiments', paper presented at NATO Advanced Summer Institute on Automatic Speech Analysis and Recognition, Bonas, France, (1981).

[5] 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).

[6] D. A. Waterman, 'A Guide to Expert Systems', Addison Wesley, 152-156, (1986).

[7] P. Ladefoged, 'Preliminaries to Linguistic Phonetics', University of Chicago Press, (1971).

[8] P. Ladefoged, J. De Clerk, M. Lindau and G. Papcun, 'An auditory-motor theory of speech production', Working Papers in Phonetics, Vol. 22, 48-75, (1972).

[9] M. Rossi, 'Niveaux de l'analyse phonetique: nature et structuration des indices et des traits', Speech Communication, Vol. 2, nos. 2-3 Special Issue, 91-106, (1983).

[10] R. Jakobson, G. Fant and M. Halle, 'Preliminaries to Speech Analysis', M.I.T., (1951).

[11] N. Chomsky and M. Halle, 'The Sound Pattern of English', Harper and Row, (1968).

[12] P. Ladefoged, 'A Course in Phonetics', Second Edition, Harcourt Brace Jovanovich, (1982).

[13] S. Singh, 'Distinctive Features: Theory and Validation', University Park Press, Baltimore, (1976).

[14] S. Singh and G. Woods, 'Perceptual structure of 12 American vowels', J.A.S.A., Vol 49, 1861-1866, (1971).

[15] P. D. Green and A. R. Wood, 'A representational approach to knowledge-based acoustic-phonetic processing in speech recognition', Proc I.C.A.S.S.P., paper 23.4, (1986).

[16] T. M. Neary, 'Phonetic feature systems for vowels', I.U.L.C., Bloomington, (1978).

[17] C. H. Coker and O. Fujimura, 'Model for specification of the vocal tract area function', J.A.S.A., Vol. 40, 112-123, (1966).

[18] P. Mermelstein, 'Articulatory model for the study of speech production', J.A.S.A., Vol. 53, 1070-1082, (1973).

[19] J. P. Zerling, 'Articulation et coarticulation dans les groupes occlusive-voyelle en francais', These de 3e Cycle, Nancy, (1979).