Studies of perceptual confusions reinterpreted as evidence for non-linear phonology[*]

WON CHOO & MARK HUCKVALE

1.0 Introduction

Human cognitive processing appears to be able to deliver a segmented phonological description of words from a continuous speech signal generated by a smoothly changing vocal tract. Aspects of this activity are generally accepted: that the peripheral auditory system is able to extract a range of spectral and temporal features of the signal; that there exists a lexicon of morphological units which may be identified by phonetic strings. But what fits in the gap between these two is still moot.

We use the term phonetic processing to cover this gap between what the auditory system delivers and what the lexicon requires; and in this paper we ask some questions about how it might be constructed and whether one kind of psychological testing can shed some light on opposing hypotheses.

1.1 Conventional view of phonetic processing

The conventional view is that between auditory patterns and phonological units there exists a layer of phonetic features characteristics of the signal which are sensitive in some sense to the contrasts required by the lexicon. A low-level cognitive process of feature extraction is developed by the child through which aspects of the signal relevant to distinguishing words (and intonational/paralinguistic patterns) are emphasised, while irrelevant aspects are suppressed.

Historically, there has been an influential interpretation of these features: that they represent details of the speaker's articulation (or intended articulation) of the utterance. The recovery of articulatory features from the signal is developed by listening to one's own speech - and indeed one can think of the feature extraction level as normalising others' speech into the 'equivalent' articulations of one's own vocal apparatus. The economy of this, the Motor Theory of speech perception (Liberman & Mattingly, 1985), has lead to its influence, despite it remaining cognitively mysterious.

[*] We are grateful to John Harris for his helpful comments on an earlier draft.

A second view of feature extraction puts more weight on aspects of auditory patterning - that certain combinations of auditory pattern elements 'trigger' a phonetic feature detector. In this view the auditory patterning of speech sounds is paramount, and we learn to articulate to generate effective perceptual features. One influential view is that phonological contrasts have an invariant auditory distinctiveness; that, say, spectral shape just after a stop burst is sufficient to identify (phonological) place (Stevens & Blumstein, 1979).

From this feature level - a description of the signal in which its special phonetic character is enhanced - there must then be a second stage of processing which recovers an underlying lexical pronunciation. This stage - let's call it phonological processing - must segment the speech in time, must 'undo' coarticulations, assimilations, elisions and allophonic variations to produce a representation adequate for lexical access. For the conventional view, this stage remains the most awkward since conventional models of production use phonological structures (linear segments, re-write rules) that simply don't work backwards. Because production rules lose phonological information and smear information in time along the string, there are always multiple backward interpretations.

This kind of argument leads to a second view of phonetic processing.

1.2 Non-linear phonetic processing

Many problems of recovering a segmented phonological representation from a continuous auditory representation can be neatly finessed by substituting a non-segmented phonological representation. Taking the syllable as a phonological structure rather than phonemes allows acoustic information about lexical identity to be spread within the syllable to no disadvantage to distinctiveness; that /bad/ has no acoustic portions uniquely /b/ or /d/ is irrelevant, we simply interpret features of the syllable as evidence for the phonological structure of the syllable.

The substitution of different phonological structures need not affect the conventional view of feature extraction: there still remain alternative views about whether the features represent articulatory or auditory aspects of the signal; we have only substituted a different phonological processing stage.

However, this view hasin turn come under attack from phonological theories which attempt to unify the two stages of phonetic processing. Instead of a stage which produces continuously valued features continuous in time which require phonological interpretation; the view of Government Phonology (Kaye et al. 1985, 1990) is that features can be extracted from the signal which are themselves phonological and suitable for lexical access. This is a kind of radical extension of the 'invariant acoustic cues' hypothesis, in which spectral shape, say, is the phonological feature for place. Univalent 'elements' in this theory are just the outputs of a set of feature detectors operating on the auditory output. The presence or absence of elements leads directly to a phonological interpretation without any rule-governed inferencing.[1]

1.3 Interactionist view of phonetic processing

A third model of phonetic processing does not view the lexicon as a separate, passive structure waiting to receive phonetic input from below. In models such as TRACE (McClelland & Elman, 1986) and the Network Lexicon (Huckvale, 1990), the feature representation is used to activate words directly, which then compete in the degree to which they 'explain' the input. This competition is mediated by phonological structures which identify phonologically equivalent units across lexical entries. TRACE takes a conventional linear segmented phonology, whereas the Network Lexicon allows for a variety (and even a mixture) of phonological systems.

What is appealing about these models is that they provide a usage of phonological knowledge which matches the definitions of phonological units. We can both implement /b/ and define /b/ by linking together all words that contain /b/. In contrast, the models above need some phonetic definition of /b/ separate from the words in the lexicon.

1.4 Making choices

Given such uncertainty about the nature of phonetic processing, one is justified in asking how competing models should be judged. There are different kinds of evidence that may be used:

- Categorical perception studies
- Studies of production errors
- Lexical access studies
- Attempts at machine recognition of speech
- Acoustic-Phonetic studies
- Psychological studies of perceptual distances

[1] How the continuous auditory representation is segmented in time is still unclear.

Studies of categorical perception relate phonological choice to acoustic structure to determine which aspects of acoustic patterning are responsible for decisions: i.e. what acoustic material the feature detectors operate on. The general finding is that there are many interacting cues to any phonological distinction. These studies also use artificial stimuli under artificial test conditions.

Production errors show a strong segmental influence - substitutions involve linear phonological units - but then substitutions also tend to obey legal phonotactic sequences, implying the influence of the lexicon.

Lexical access and shadowing studies show that listeners are able to identify words given only the first part of the phonetic evidence. The quantity of evidence required approximates an integral number of linear segments sufficient to cut down the number of lexical choices to one.

Machine recognition systems have only been able to demonstrate bottom-up phonemic transcription performance of about 70%. This only serves to emphasise the importance of high level knowledge.

Studies of production - identifying the characteristics shared by all given examples of a phonologial units are pre-disposed to give optimistic results of invariance. It is always possible to design a feature detector that gives a 100% hit rate at detecting a given phonological feature of the sound stream. Unfortunately that information is useless unless the detector also has a very low false-alarm rate. We must continue to emphasise the distinctiveness of patterns, not just their identity.

All these studies contribute some information to a debate about the nature of phonological processing, but they all bring with them a quantity of phonological prejudice which makes them unreliable indicators. Categorical perception studies use linear segments and specify the acoustic patterns to test; production errors have to be recorded as phonetic transcription; lexical access studies subjects reply with words and their reaction times are converted to transcription units. ASR uses linear phonological units themselves as acoustic models. Acoustic-phonetic studies are not reliable separately from recognition.

An interesting alternative to these then are psychological studies which allow a more unbiased interpretation of human speech perception performance. Starting with CV syllables demonstrating known contrasts, information about perceptual similarity is converted to an analysis of the decision making process independently of phonological system or proposed acoustic-phonetic features.

2 Perceptual studies of phonetic features

2.1 Design of perceptual studies for consonants

Most perceptual studies of consonants are concerned with determining the number and nature of the perceptual features utilised in the identification and internal representation of consonants. Investigations into the nature of these perceptual features fall into two classes depending on whether:

1) Features or feature systems are proposed in advance of analysis to interpret perceptual responses or;
2) Features are empirically determined from perceptual confusions by methods such as multidimensional scaling (MDS).

The perceptual tests are conducted with degraded speech material, utilising noise masking, selective filtering, segment deletion and peak clipping, or through the use of cross-linguistic settings or phonetic context conditioning. The stimuli are typically consonant-vowel (CV) syllables which differ only in the identity of their constituent consonant phonemes. Perceptual responses are elicited using various psychological methods such as identification, recall of speech sounds in short-term memory, or similarity judgement of pairs and triads of speech stimuli. The results are tabulated either in the form of confusion matrices or distance matrices compiled from subjects' reactions to the stimuli.

2.2 Early studies of consonant perception

From the perceptual confusion/similarity matrices resulting from such experiments, there then follows an analysis based on the specification of a set of hypothetical perceptual features upon which information is transmitted by a determination of the relative importance of these features (see Table 1). In this type of analysis, the evidence for a particular feature is indicated by a high level of utility for the feature. An important result of these experiments is that the relative importance of the features varies according to the experimental conditions.

Studies	Condition	Language	Response	Context	Analysis	Features in the order of importance
Miller & Nicely (1955)	noise frequency distortion	English	confusion	CV	covariance	nasal, voice, duration, frication, place
Singh & Black (1966)	noise freq. distort. temporal segmentation	English Hindi Arabic Japanese	error	VCV	analysisi of variance	nasal, place, liquid, voice, duration, frication, aspiration
Singh(1966)	freq. distort. temporal segmentation	English Hindi	error	CV	rank correlation	voice, place (for English) place, voice (for Hindi)
Ahmed & Agrawal(1969)	quiet	Hindi	identification	CV VC	rank correlation	affricate, nasal, aspiration, frication, liquid, voice, continuant, place, (CV) affricate, flapped-liquid, frication, liquid, place, voice, nasal, continuant, aspiration (VC)
Gupta, Agrawal & Ahmed (1969)	peak clipped signals	Hindi	identification	CV VC	rank correlation	affricate, frication, aspiration, voice, nasal, liquid, continuant, place (CV) affricate, flapped-liquid, liquid, place, voice, frication, continuant, nasal, aspiration (VC)

TABLE 1

Summary of perceptual features in the order of their statistical significance

Study	Test	Languages	Response	Context	Feature systems compared in order of preference
Wickelgren (1966)	short-term memory	English	% of correct feature recall	CV	1. Wickelgren 2. Miller&Nicely 3. Halle
Singh(1970b)	signal/noise condition	English (English & Hindi subjects)	triadic comparison	initial & final syllable positions (CVC)	1. Miller & Nicely 2. Halle 3. Wickelgren
Wang & Bilger (1973)	noise	English	identification	CV VC	1. Miller & Nicely (1955) 2. Singh & Black (1966) 3. Wickelgren(1966) 4. Chomsky & Halle (1968) 5. Singh, Woods & Becker (1972)

TABLE 2

Summary of consonant perception in terms of different feature systems

Feature systems;

Wickelgren: voicing, nasality, openess, place
Miller & Nicely: voicing, nasality, affrication, duration, place
Halle: vocalic, consonantal, grave, diffuse, strident, nasal, continuant, voiced
Singh & Black: voicing, nasality, friction, place, duration, liquid
Chomsky & Halle: vocalic, consonantal, high, back, low, anterior, coronal, voice, continuant, nasal, strident
Singh, Woods & Halle: place, nasal, sibilant, voice, plosive

Study	Condition	Response	context	analysis	Na.	Vo.	Sib.	Cont	Pla.	Son
Wilson(1963)	noise low-pass high-pass	open choice	CV	MDS (S,W)	√	√	(√)	(√)		NA
Johnson (1967)	noise low-pass high-pass	open choice	CV	HCS	√	√	(√)			
Shepard (1972)	noise low-pass high-pass	open choice	CV	MDS (S)	√	√	(√)	√		NA
Wish (1970)	noise low-pass high-pass	open choice	CV	INDSCAL	√	√	√	√	2nd form. tran.	NA
Singh, Woods & Becker (1972)	quiet	ABX scaling ME	CV	INDSCAL	√	√	√		√	√
Mitchell & Singh (1974)	noise quiet	ABX	CV in a sentence	INDSCAL	√	√	√	√	√	NA
Soli & Arabie (1979)	noise freq. distort.	confusion	CV	INDSCAL	periodicity/burst, F1, F2, spectral dispersion					

TABLE 3

Summary of perceptual features interpreted for MDS dimensions

Na., nasality; Vo., voicing; Sib., sibilancy; Cont., continuancy; Pla., place; Son., sonorancy.
NA means that relevant features are not included in the stimuli.

An alternative approach is to compare feature systems as a whole to predict the psychological results, rather than use individual features within the system. A summary is presented in Table 2. For example, Wickelgren (1966) uses three articulatory feature systems to predict the relative frequency of consonant confusions in short-term memory, and compares the systems in terms of the percentage of prediction confirmed by the data. The Wickelgren feature system is shown to be a better predictor of the short-term memory confusions than the Halle or the Miller & Nicely feature system. Using the same feature systems but with similarities rather than confusions, Singh (1970b) found a different result in that Wickelgren's system was less effective in predicting the perceptual errors than the other two systems. The difference was attributed to the different data collection methods. Wang and Bilger (1973) provide an assessment of diverse feature systems and conclude that:

> ... for most confusion matrices several feature system can be shown to account equally well for transmitted information, and that across syllable sets and listening conditions, there is little consistency in the identification of perceptually important features.

Thus for these early studies, the statistical effectiveness of a feature set was found to be dependent on the contextual effects of the particular stimuli and the listening conditions employed. As Table 2 shows, there seemed to be no single feature system that was "best" for describing perceptual relationships between speech sounds.

2.3 MDS studies for consonants

The second type of consonant perception study involves the "extraction", as opposed to hypothesis, of a set of perceptually distinct features determined empirically by methods such as **multidimensional scaling (MDS).**

MDS provides a means of constructing spatial representation of the judgements of a listener to a perceptual task whereby the feature analysis underlying a set of perceptual responses may be represented as a multidimensional psychological/ phonological space. In this representation, the perceptual distances within a set of objects are reflected in the spatial separation between the objects in the space. That is, MDS places the stimuli in an n-dimensional space such that the distance between the objects in this space corresponds to the empirically obtained distances estimated in a perceptual experiment. The importance of the MDS technique is that only a few dimensions are required to model a large number of perceptual judgements. They may then be normally interpretable as the distinctive dimensions of perceptual analysis.

The perceptual implications that can be drawn from an MDS analysis are, however, dependent on the type of scaling method used. This is because the standard MDS analysis does not provide an orientation of the axes to the solution. Therefore, other knowledge is necessary to guide a rotation of the coordinate axes to permit an interpretation of the dimensions. Any conclusions drawn are then susceptible to the criticism that some other alternative interpretation would have been equally adequate for the data according to the rotation performed. This is shown in the first three studies listed in Table 3 that involve reanalyses of the Miller and Nicely (1955) consonant confusion data by different MDS procedures. Wilson (1963) used an earlier version of Shepard's (1962) MDS technique and his own adaptation of that technique. Johnson (1967) developed a hierarchical clustering scheme. This method utilises perceived distances between stimuli and converts them into a series of rank-ordered diameters. Shepard (1972) used his own MDS technique. In all three cases, the features nasality and voicing are interpreted as corresponding to the first two dimensions. There is also some argument over two further dimensions of sibilance and continuance; represented by ( ) in Table 3.

There is a more advanced MDS technique called individual differences scaling (INDSCAL) which calculates an orientation of the axes to the solution which cannot be changed without worsening the overall fit to the perceptual data. With INDSCAL the claim that the dimensions have a perceptual reality is, therefore, strengthened. INDSCAL considers perceptual strategies across individual subjects and determines the relative salience or weight of each dimension for each subject. INDSCAL does, however, suppose that individuals share the same perceptual dimensions when making judgements on a common set of stimuli, and assumes that subjects differ only in terms of the weight, or salience, that they attach to each dimension. Under normal conditions, however, with well-defined perceptual tasks, it is assumed that there is a best solution for the dimensions which accounts for the individual variance. It has been conjectured that INDSCAL dimensions have perceptual reality and studies in several areas of human perception have shown that INDSCAL results have corresponded to previously established models of perceptual processes (Wish and Carroll, 1974).

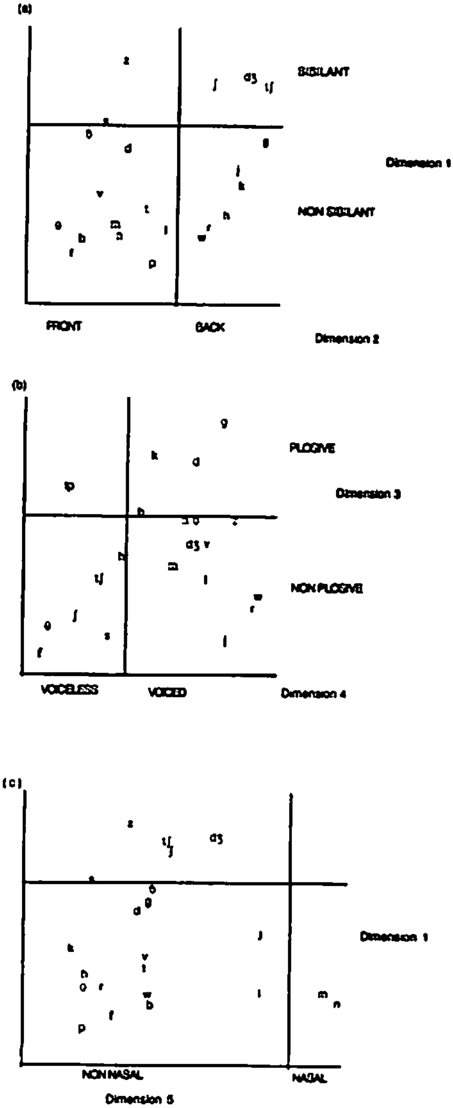
In the area of speech perception, a direct comparison can be made of the INDSCAL analysis of the Miller and Nicely (1955) data in Wish (1970) with those of the MDS analyses of the same data listed in Table 3. Wish reported a five dimensional solution to be optimal, as opposed to the three-feature dimensions of earlier work, corresponding to the features nasality, voicing, sibilance, continuance, and the second-formant transition; here the INDSCAL technique shows it can recover more perceptually salient features. INDSCAL does not provide an explicit measure for selecting the degree of dimensionality, although in general the lower dimensionality divides the stimuli in broad phonetic categories while the increased dimensionality provides a finer set of categories. For example, a two dimensional result may be interpreted as voicing and nasality. If extended to five dimensions, the voicing may show further categories of voiced and voiceless stops (e.g. dimensions 3 and 4 in Figure 1b).

Table 3 presents a summary of a number of MDS studies showing a variety of data collection methods, different methods of eliciting responses, different phonetic contexts and the different MDS methods. Despite these variations, the perceptual dimensions obtained are remarkably stable and in general, correspond to five features of the consonants:

- nasality
- voicing
- sibilance
- place of articulation (front/back)
- sonorance (plosive/nonplosive)

Figures 1a,b,c show two-dimensional plots of these results - in each figure, a horizontal or vertical line is interpretively positioned for binary separation of the consonants (Singh, 1976). Figure 1a shows dimensions 1 and 2 corresponding to sibilance and place of articulation which separates the consonants articulated back of the alveolar ridge from those articulated against or in front of the alveolar ridge. Figure 1b shows dimensions 3 and 4, separating the plosives from nonplosives and voiced consonants from the voiceless consonants (except for /k/). The separate clustering of the consonants /w r l j/, in dimension 4 can be interpreted as an extra perceptual dimension of INDSCAL analysis. Figure 1c shows the plotting of the sibilant dimension 1 against dimension 5 which is the nasality dimension.

Figure 1. Five-dimensional INDSCAL configurations for consonant perception (from Singh, Woods & Becker, 1972).



Study	Stimuli	Context	Response	Analysis	Perceptual Dimensions
Hanson (67)	natural & synthesized	none	similarity judgment	factor analysis	advancement, height (F2) (F1)
Pols et al (69)	i y ø e ɛ œ a ɑ ɔ o u	h-t	triadic comparison	MDS	advancement, height
Singh & Woods (70)	i I e ɛ æ ɑ ɔ o U u ʌ ə	none	dissimilarity judgment	MDS	advancement, height, retroflex
Anglin (71)	i I e ɛ æ ɑ ɔ o U u ʌ ə	h-d	dissimilarity judgment	MDS	advancement, height, retroflex, tenseness
Shepard (72)	i I e ɛ æ ɑ ɔ o U u	h-d	identification (conf.data)	Exponential analysis	advancement(F2), height(F1), tenseness
Fox (74)	i eI ɑ æ aI o u	-t	dissimilarity	INDSCAL	height, front onset, back/front glide
Terbeek (77)	i i y ø e ɛ æ o ʌ ə ɑ ɑ	bəb-	triadic comparison	INDSCAL	roundness, height, retroflex, advancement, mid/nonmid
Fox (78, 82)	i I ɛ æ ɑ ɔ U u	h-d	dissimilarity	INDSCAL	advancement, height, round
Fox (83)	i I eI ɛ æ ɑ ʌ ɔ oU U u ɔI aI aU ju	h-d	dissimilarity	INDSCAL	advancement, height, low-back onset, mid/nonmid
Fox (85)	i I ɛ æ a U u	h-d	scaling	INDSCAL	advancement, height
Rakerd & Verbrugge (85)	i I ɛ æ ɑ ɔ o U u ʌ	hə'd-də V	triadic comparison	INDSCAL	advancement, height, tenseness
Fox & Trudeau (88)	i I ɛ æ ɑ ɔ oU U u ʌ ə (Esophageal vowels)	h-d	scaling	INDSCAL	advancement(F3-F2), height(F1-F0), retroflex(F3)

TABLE 4

Summary of vowel perceptual dimensions elicited by MDS

2.4 MDS results for vowels

Table 4 is a summary of studies involving the perception of vowels (based on Fox, 1983), most of which use MDS analyses. Included are the stimulus vowels used (the majority of which are English) phonetic context, response type, analytic technique, and the dimension labels. As can be seen, the perceptual dimensions extracted are remarkably consistent and related either to the articulatory features or acoustic parameters of the stimuli. The most common feature interpretation of the dimensions is in terms of the articulatory features: height and advancement. Additional dimensions show differing results for instance, Singh and Woods (1971) found a retroflexion dimension; Fox (1982) obtained a rounding dimension while in Rakerd and Verbrugge (1985) and Anglin (1971), the third dimension corresponded to tenseness. The obtained feature differences are generally attributed to different stimulus sets and phonetic context used; the retroflexion feature can of course be only retrieved in rhotic vowel systems.

In terms of acoustic properties of stimuli, the first two dimensions are shown to be most highly correlated with F1 and F2 (Shepard, 1972), although it is suggested that duration and dynamic properties may act as supplementary cues to vowel identifications (Fox, 1983).

2.5 Acoustic interpretation of the perceptual dimensions

We have emphasised thus far a traditional articulatory feature interpretation of the dimensions. But the perceptual data could equally be described in auditory features such as spectral shapes and transitions. The studies which have such orientation are briefly discussed below.

In Fox (1983), acoustic parameters relevant to vowel perception of English monophthongs and diphthongs are assessed by multiple linear regression, giving five dynamic and steady state measures of the first three vowel formants. There are two major assumptions of his acoustic interpretation: that perceptual features are best interpreted as the integration of several acoustic cues (i.e. multiple regression technique is used to analyse the relationship between each perceptual dimension and all the various formant measures); and that the dynamic formant structure of diphthongs may contribute to the perceptual processing. The general conclusion is that the first two perceptual dimensions are mainly explicable in combination of the first two formant frequencies, with some additional durational effect, and the third dimension by the combination of F2-F1 transition and changes in F2, without having recourse to the dynamic acoustic information.

The acoustic dimensions found in Rakerd and Verbrugge (1985) for isolated vowels closely resemble the articulatory interpretations; advancement dimension correlated with F2 and F3, height with F1 and tenseness with duration. In an experiment involving vowels in context, a coarticulatory effect is demonstrated by a lower correlation value between formants and perceptual dimensions; here tenseness is claimed to be related to offglide proportion rather than the vowel duration itself.

There have been fewer acoustic interpretations of consonant perceptual dimensions in the literature, and even those are limited to the consonants in CV context only. For the consonant confusion data of Miller and Nicely (1955) reanalysed by INDSCAL (Soli and Arabi, 1979), the accounted acoustic properties of the speech signal include temporal relationships of periodicity and burst onset, shape of voiced F1 transition, shape of voiced F2 transition and amount of initial spectral dispersion.

An issue now is whether the acoustic/auditory properties or the articulatory features are to be preferred to explain perceptual dimensions. The only work that sheds some light is that of Fox (1985) which specifically addresses the question of 'the degree to which perceptual similarity judgments (and the perceptual dimensions obtained after such judgments have been analysed using MDS techniques) are sensitive to the acoustic nature of the stimuli being compared'. Fox investigates the relationship between perceptual dimensions and subphonemic acoustic information (such as vowel formant values) with two sets of vowel stimuli, /i, ɪ, ɛ, æ, a, ʌ, u/ differing only in the formant values of the subset /ɪ, ɛ, æ/. Then, the values in the perceptual distance matrix for each set are compared, firstly by multivariate analysis of variance, then by INDSCAL. Fox found the difference in acoustic distance between the vowels /ɪ, ɛ, æ/ for the two separate sets was directly reflected in the difference in perceptual distance. Because the vowels /ɪ, ɛ, æ/ are unstable across most American dialects and idiolects, the experiment is repeated with acoustic distance variations in [i]-[ɪ] and [u]-[ʊ] pairs. The result demonstrates that subject's similarity responses are influenced by the phonetically uncategorized acoustic domains apart from their phonetic labels.

3 Non-linear phonetic processing

We are now in a position to return to the opening discussion. Having established good evidence that traditional articulatory-based features fit the psychological feature space and some slight preference for an acoustic interpretation of those features, we can ask; do non-linear phonological features fit the data any better?

We start with a brief description of the low-level structure of Government Phonology and then re-interpret the MDS results.

3.1 Elements

The elements of Government Phonology are the basic building blocks of language, the lowest rank in phonological hierarchy (Kaye et al. 1985, 1990). The key notions of elements are:

1) universality
2) autosegmental tier representation
3) univalency
4) head-dependent relation
5) autonomous phonetic interpretation

Universality means that the elements are a part of common linguistic competence shared by all human languages. Regardless of whether a particular language utilises all of the elements or a subset of them, they are assumed to be sufficient to characterise all observed phonological processes. Each element is represented on its own **autosegmental tier**, reflecting the fact that the element operates independently of all other elements in phonological processes. The autosegmental tier representation forms part of a two-dimensional melodic grid; columns correspond to timing and rows correspond to the element tiers. At each grid point, an element is **univalent**; i.e. present or absent rather than plus or minus. Thus, each melodic unit (conventionally a phoneme) can be represented as a composition of elements attached to the relevant timing slot (see the configurations in Tables 5 and 6). **Head-dependency** describes the element 'permutations', that is, the way in which the elements are arranged together to form more complex segments. For example, when the vowel elements, A and I are combined there are two permutation possibilities, depending on which element is the head, and which element is the dependent (the head is underlined)

I + A = [e]   I + A = [æ]

Obviously, as more elements are joined together the permutation possibilities increase, which will enable the description of the full range of vowels and consonants in any particular language. Lastly, each element is claimed to be a **phonetically autonomous** entity which is pronounceable on its own, as a head of a simple segment. For example, the vowel u is made up of one element U (see

Table 5). This claim receives empirical support from the acoustic analyses of elements in works by Lindsey and Harris (1990), Harris and Lindsey (1992) and Williams and Brockhaus (1992).

Of the five claims for element theory, the most important to us here are 'univalency' and 'autonomous phonetic interpretation', the combination of which leads to invariant acoustic forms for elements.

3.2 Acoustic invariance of elements

In Government Phonology, the acoustic property of each element is defined in terms of unique patterns in the spectrogram.

Vowels are made up of three major elements A, I, U which have supposed acoustic specification (see Figure 2):

A 'Mass': energy higher in middle (~100-1600 Hz) band then at top and bottom. Convergence of F1 and F2.

I 'Dip': energy lower in middle (~900-2000 Hz) than either side. Convergence of F2 and F3.

U 'Rump': energy below middle (~50-900 Hz). Convergence of F1 and F2.

Two additional vowel elements are:

@ 'Neutral': no salient acoustic property.

% 'ATR': accentuation of spectral shape associated with A, I or U.

The elements used for the description of consonant manner are:

h 'Noise': high-frequency aperiodic energy; the release burst and subsequent noise corresponds to aspirated quality.

? 'Stop': abrupt decrease in overall amplitude; a brief period with little energy on either spectrum or waveform; this period varies however, according to the type of segment.

R 'Coronal': formant transitions associated with "coronality".

Figure 2a. Elemental patterns of simplex vowels, [a, i, u] (Harris and Lindsey, 1992).

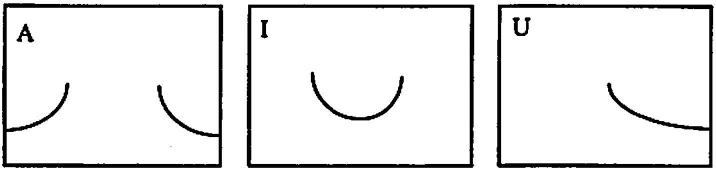


Figure 2b. Compound vowel element patterns, [e, o].

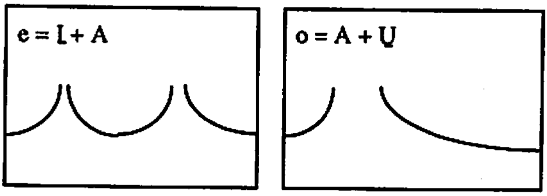
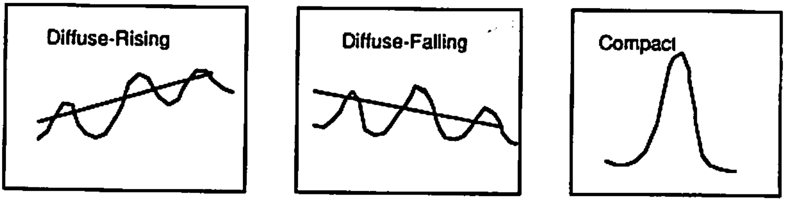


Figure 3. Supposed acoustic specifications of U (diffuse-falling), R (diffuse-rising) and I/@ (compact).



The elements used for the description of consonant place are (Stevens & Blumstein, 1979; Blumstein, 1986) (see Figure 3):

U Diffuse-falling; flat distribution of spectral energy or a relatively sustained spectral energy in the low frequencies, around 1500 Hz (labial consonants).

R Diffuse-rising; greater or relative sustained energy at the high frequencies than diffuse-falling (alveolar consonants).

I Compact; same as in velars but with higher frequency peaks (palatal consonants).

@ Compact; one or two peaks dominating the spectrum in the frequency regions between 1200 - 3500 Hz (velar consonants).

The elements used for the description of nasality and voicing are:

N 'Nasal': broad resonant peak at lower end of frequency range.

L 'Low tone': marked drop in F0 relative to that of an adjacent vowel.

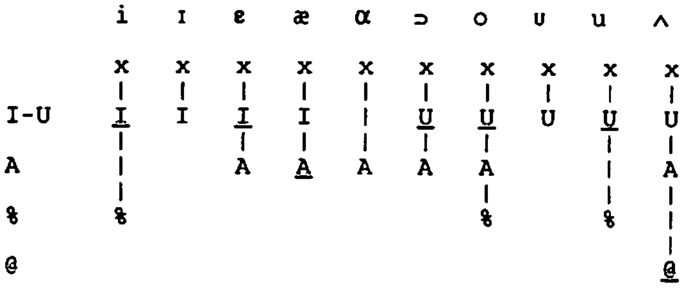
H 'High tone': raised fundamental frequency high-frequency aperiodic energy.

(For more details of the above descriptions see Lindsey and Harris (1990), Harris and Lindsey (1992) and Williams & Brockhaus (1992).)

3.3 Reinterpretation of MDS data for vowel elements

Table 5 shows 10 American English vowels, expressed in terms of their compositional elements (heads underlined). Each tier is dedicated to a particular element except for the palatal/ labial tier. This is because the elements I and U do not behave independently of each other in English. One element is present exclusively of the other; thus the vowel /y/ does not occur.

Table 5. Elemental representation of ten American English vowels.



	i	ɪ	ɛ	æ	ɑ	ɔ	o	ʊ	u	ʌ
	x	x	x	x	x	x	x	x	x	x
I-U	I	I	I	I	|	U	U	U	U	U
A	|		A	A	A	A	A		|	A
%	%						%		%	|
@										@

Rakerd (1983) obtained numerical ratings of similarity using these ten vowels and analysed them by means of individual differences scaling. The percentage of variance in the perception data indicated a three-dimensional configuration. The group space for all subjects (isolated-vowels and consonantal-context conditions combined) is shown in Figure 4. It can be seen that Dimension 1 matches well to elements I-U; this distinguishes such vowels as /i, ɪ, ɛ, æ/ (which are projected onto the "I" end of Dimension 1) from such vowels as /ʌ, ɔ, o, ʊ, u/ (which project onto the "U" end of dimension 1). The head-dependent relation is not interpretable in dimension 1; otherwise, the projections of the vowels /ɛ, æ, ʌ, ɔ, o/, which are distinguishable by the headedness of I and U elements in composition, must be positioned accordingly. For example, /ɛ/ should be positioned nearer to the "I" end than /æ/. The vowel /ɑ/, which is composed of unitary element A, is projected onto the middle of "I-U" dimension (absence of elements, I, U). Dimension 2 matches to A ("Mass" pattern), separating vowels in terms of degree of A-ness present: A is the head element for the vowels /æ, ɑ, ɔ/, A is the dependent for vowels /ɛ, ʌ, o/, and A is absent for the vowels /i, ɪ, u, ʊ/. The vowel /o/ is an exception to this interpretation because the A element is the dependent. Dimension 3 approximates to the ATR element; ATR is never manifested as a head element in the vowels. This distinguishes the vowel pairs /i, ɪ/, /u, ʊ/ and /o, ɔ/. A fourth dimension is required by the element system, but not needed in this interpretation, which corresponds to the @ element, used to separate /ʌ/ from all the other vowels. (The element, cold-vowel @, is not assigned to a specific acoustic property and acts as a default vowel in the system.)

Figure 4. Perceptual structure of vowels reinterpreted in terms of elements (data from Rakerd, 1984).

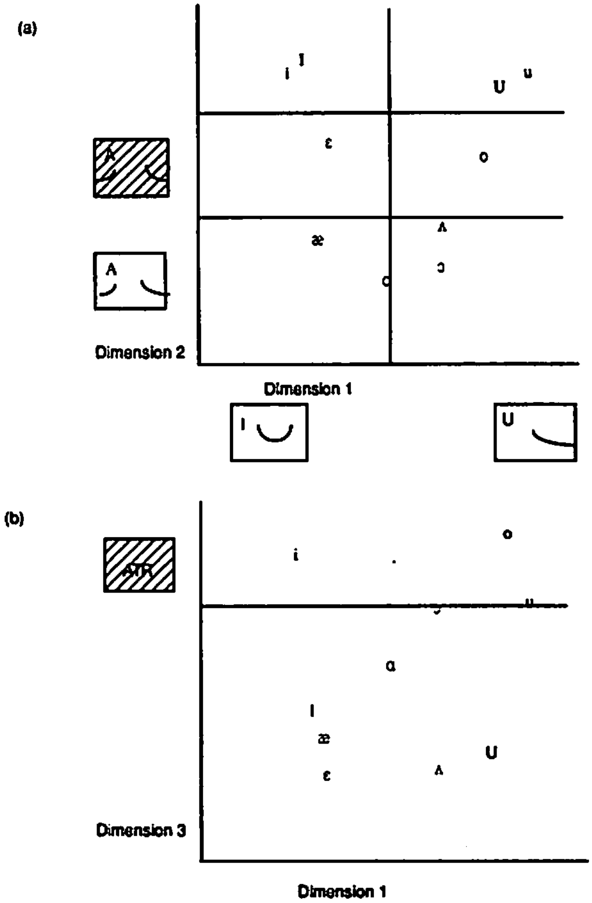
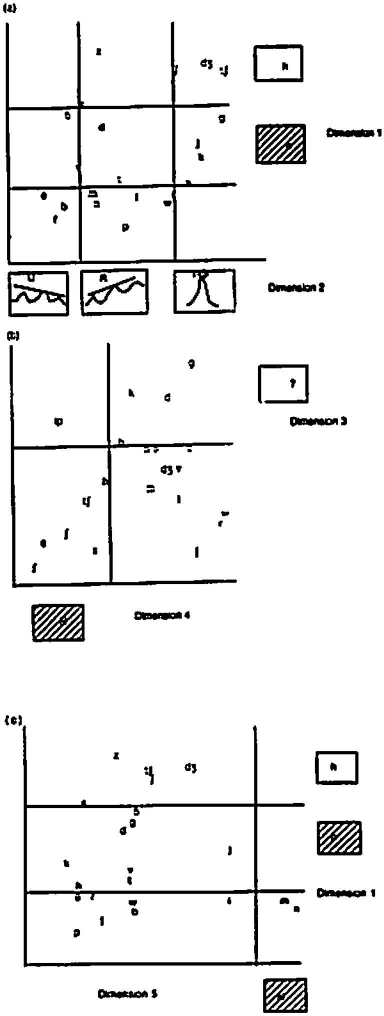


Figure 5. Perceptual structure of consonants reinterpreted in terms of elements (data from Singh, Woods & Becker, 1972).



3.4 Reinterpretation of MDS data for consonant elements

The element representations of English voiceless consonants in pre-vocalic positions are given in Table 6. The corresponding voiced consonants lack the high tone element, H, which is manifested as 'aspiration' in prevocalic positions. There is an ambiguous representation for the nasals in that there appears to be insufficient evidence to determine whether the element ? or N acts as the head.

Table 6. Elemental representation of English consonants.

	p	t	k	f	θ	s	ʃ	tʃ	l	r	m	n	j	w	h
	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
?	?	?	?					?	?		?	?			
R		R			R	R		R	R	R		R			
h	h	h	h	h	h	h	h	h							h
U-I-@	U		@	U			I	I		@	U		I	U	
H	H	H	H	H	H	H	H	H							
N											N	N			

The most comprehensive MDS study on consonant perception is by Singh, Woods and Becker (1972), on which our analysis is based. It involves three data collection methods, equal appearing interval scale, magnitude estimation and triadic judgements, a large number of subjects and 22 initial consonants. The perceptual data are analysed using both MDS and INDSCAL.

The configurations of the 22 consonants in the five-dimensional INDSCAL stimulus space are given in Figure 5 (a) (b) and (c). (The horizontal and vertical lines in these figures have been interpretively positioned to segregate the consonants.) Three groups of phonemes project onto the first dimension (Figure 5a) in the following order; /s z ʃ d tʃ/, /v δ d t k g j h/ and /θ f b p m n l w r/. The composition and arrangement of these three groups corresponds approximately to an ordering of the consonants on the dimension according to the degree of presence of the h element. Headedness of the h element in a segment is indicated by an unshaded box whereas dependency is indicated by a shaded box. Exceptions are /b p f θ j h/: /f/ and /h/ should be in the unshaded h-region where h manifests as a head element, /b, p, θ/ should be placed in the shaded-h region with dependent h element, and /j/ should be placed with the segments with no h element in composition.

The projection of the consonants on the second dimension of Figure 5a also shows three groups. In the first group are the labials /b f v/ and dentals /θ ð/. The alveolars, /d t s z n l r/ form the second group, and in the third group are palatals /j tʃ dž/ and velars /g k h/. The corresponding acoustic properties are specified in terms of the elemental patterns on the figure. Palatals and velars share the same 'compactness' pattern (section 3.2, Blumstein, 1986). The segments /m p w/ are, however, apparent exceptions to the R-element category.

Dimension 3 separates the consonants, /b p d t g k/ from the rest of the consonants, according to presence versus absence of ? element in the composition. Headedness of the element ? is not attested in the perceptual data. Segments /m n/ and /l/ are the exceptions to this interpretation.

Dimension 4 matches the H element. As mentioned above, voiceless consonants in prevocalic positions contain an H element, of which the acoustic manifestation is 'aspiration'. In English, phonololgically 'voiced' obstruents are not specified in terms of any particular element. These, together with sonorants, are projected onto the lower part of Dimension 4.

The fifth dimension of Figure 5c separates the nasals /m n/ from all the other consonants corresponding to element N.

4 Discussion

The result of the multidimensional scaling analysis demonstrates that the observed perceptual relationships among the English sound segments in Rakerd (1984) and Singh, Woods and Becker (1979) experiments can be fitted into the element descriptions in Government Phonology. However, the question remains whether this non-linear phonological model of elements explains the data any better than the traditional feature systems described in section 2.

Beginning with the evaluation of the vowel I-U dimension of Figure 4a, note that two distinct groups are defined on the dimension depending on whether I or U is the head in the elemental composition for each vowel; so the vowel /ɑ/ is placed in the middle. It is interesting to note that the elements I and U, which are hypothesized to occupy the same autosegmental tier (see Table 5), are placed on the same dimension. Thus, this interpretation seems to be as valid as the traditional front/back articulatory description.

Similarly for the element A dimension in Figure 4a, the axis is labelled to characterize the three distinct grades of the element decomposition of segments: /ɑ/ which consists of unitary element A as head is placed at the bottom of the scale, and /æ/ is appropriately separated from it according to the headed or dependent manifestation of the A element.

Considering dimension 3 in Figure 4c, ATR is a feature which also exists in conventional feature systems, and the element label is not so different from the original one except that ATR element is either present as a dependent or not present at all, but never as a head, in the non-linear system. Thus the interpretation is parallel to the one using the traditional features.

Therefore the vowel element descriptions fit with the perceptual data rather well, though this isn't surprising given the fact that the vowel elements are not so different from traditional vowel features in terms of their labelling, even though they are in terms of their phonological role and acoustic characterization.

Considering next the appropriateness of the consonant elements for the perceptual dimensions in Figure 5, note that only the h dimension shows three distinct groupings reflecting the head-dependent relation. Now comparing dimension 1 in Figure 5a with that of Figure 1a: the previous nonsibilance category is now further divided into segments with dependent h element and segments with no h element. Among them 6 segments out of 17 are exceptions to the category. Thus, it appears that the traditional binary feature interpretation for the first dimension provides a more satisfactory account of the data.

The segment projections on the consonant place dimension of Figure 5a also do not clearly meet the criteria for element description; conforming to the hypothesis that the elements characterising the same dimension should occupy the same autosegmental tier, coronal element R should be represented on a separate dimension, but we find it as part of the place dimension along with elements U, I and @.

For dimensions H, ? and N, the distribution of the consonants is shown to approximate to element representation, but their head-dependent property is not always demonstrated. This can be partly explained by the phonological evidence that the H element never manifests as a head and the nasal element only manifests itself as a head element (assuming that the nasals are N-headed rather than ?-headed). As a result, these dimensions are only divided into two groups, as are the dimensions in the articulatory feature systems.

Thus overall it seems that it is more convenient to interpret the consonant perceptual data in terms of a subset of traditional features, selected for that purpose, as dimension labels. However, this is as far as it will go, in that this independent perceptual system is unable to account for any phonological variations in human speech processing; according to the conventional view of phonetic processing, cues or features extracted from the speech signal are processed or transformed into phonological representations, which in turn are matched against entries in the lexicon. Therefore, the perceptual dimension labels without reference to the higher level phonological processing give us a rather incomplete picture of phonetic processing as a whole.

On the other hand, non-linear phonetic processing which combines feature extraction and phonological processing stages, with its lexical acoustic units, such as that of the elements, satisfies both theoretical and perceptual perspectives; the number of the recoverable dimensions corresponds more closely to the number of elements suggested in Government Phonology than any of the traditional feature systems, and furthermore non-linear phonology is far superior in accounting for phonological variations. Therefore, it seems that the non-linear phonological system presents a more complete and satisfying account of phonetic processing, and the present study suggests this view has a certain perceptual validity. In addition, the Singh, Woods and Becker (1979) experiment by no means exhausts the set of possible perceptual studies, and other interpretations may be possible.

5 Conclusion

This study was motivated to assess different views of phonetic processing, defined in section 1, by means of psychoperceptual testing. In section 2 we have looked at previous perceptual experiments, of which the results are interpreted in terms of conventional phonological features. In section 3 we have explored a possibility of interpreting the same perceptual data in terms of non-linear phonological elements; the specific case under investigation was the elements of Government phonology. The reinterpretation of the data demonstrates that the observed perceptual relationships among English sound segments can successfully be described in terms of elements. Whether the element-based description of these dimensions is preferred to the traditional feature description must be confirmed on the basis of statistical evaluation of the stimulus configurations along individual dimensions.

In our continuing research programme, the next stage involves quantifying the elemental patterns in production, and the use of the quantified patterns to 'recognise' the sounds. The annotation labels for four English recordings (approximately two minutes each, taken from the first ESPRIT Project 2589 (SAM) CDROM database (Eurom.0)) are being used to extract the vowel segments. Spectra are derived from the speech wave forms by sampling at the middle of the vowels, and these spectra are matched against the prototype vowels. Each spectrum for the individual vowel is obtained by averaging all the possible occurrences of a particular vowel in the data, each smoothed by a cepstral algorithm. Those spectral curves which do not conform to the general patterns are discarded. The acoustic properties of vowel prototypes were similar to the A, I and U descriptions in section 3.2. These prototype vowel spectra will be used as bases to locate each vowel according to their quantified acoustic distances. We will be examining to what extent this matches the perceptual dimensions.

We also hope to use multidimensional scaling to attempt to verify the universality of the elements - for example, whether vowel systems containing mid-vowels (in which the elements I and U occupy separate autosegmental tiers) will be placed on separate dimensions.

References

Ahmed, R., S. S. Agrawal (1969). Significant Features in the Perception of (Hindi) consonants, *JASA* 45. 758-763.

Anglin, M. (1971). Perceptual space of English vowels in word context. Unpublished master's thesis, Howard University, Washington, D. C.

Arabie, P., S. D. Soli (1979). The interface between the Type of Regression and Methods of Collecting Proximities Data. In R. Golledge and J. N. Rayner (eds), *Multidimensional Analysis of Large Data sets*. Minneapolis: University of Minnesota Press.

Blumstein, S. E.(1986). On Acoustic Invariance in Speech. In J. S. Perkell and D. H. Klatt (eds), *Invariance and variability in Speech Processes*. MIT Lawrence Erlbaum Associates.

Carroll, J. D. & M. Wish (1974). Multidimensional perceptual models and measurement methods. In E. C. Carterete and M. P. Friedman (eds), *Handbook of Perception*. New York: Academic Press. Vol. 2. pp. 341-447.

Fox, R. A. (1985a). Multidimensional scaling and perceptual features: evidence of stimulus processing or memory prototypes? *J. Phonet.* 13. 205-217 .

Fox, R. A. (1985b). Auditory contrast and speaker quality variation in vowel perception. *JASA* 77. 1552-1559.

Fox, R. A. (1983). Perceptual structure of monophthongs and diphthongs in English. *Lang. Speech* 26. 21-60

Fox. R. A. (1982). Individual variation in the perception of vowels: Implications for a perception-production link. *Phonetica* 39. 1-22.

Fox, R. A. (1978). Individual perceptual variation and a perception/production link in vowels. *Papers from the Chicago Linguistic Society*, Chicago. 98-107.

Fox, R. A.(1974). An experiment in cross-dialect vowel perception. *Papers from the 10th Meeting of the Chicago Linguistic Society*, Chicago. 178-185.

Fox, R. A., Trudeau, M. D. (1988). A Multidimensional Scaling Study of Esophageal Vowels *Phonetica* 45. 30-42.

Goldstein, L. (1971). Three studies in speech perception: features, relative salience, and bias. *UCLA Working Papers Phonet.* 39. 1-87.

Gupta, J. P., S. Agrawal, and R. Ahmed (1969). Perception of (Hindi) consonants in clipped speech. *JASA* 45. 770-773.

Hanson, G. (1967). Dimensions in speech sound perception: An experimental study of vowel perception. *Ericsson Tech.* 23. 3-175.

Harris, J., G. Lindsey. (to appear). Segmental decomposition and the signal. To appear in *Phonologica* 7.

Huckvale, M.A. (1990). The Network Lexicon: A Novel Application of Phonological Knowledge in ASR. *Speech Hearing and Language: Work in Progress* UCL vol. 4. 181-194.

Jacobson, R., C.G.M. Fant, M. Halle (1963). *Preliminaries to speech Analysis: The Distinctive Features and their Correlates.* MIT Press, Cambridge, MA.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika* 32. 241-254.

Kaye, J. D., J. Lowenstamm and J. Vergnaud (1985). The internal structure of phonological elements: a theory of charm and government. *Phonology Yearbook* 2. 305-328.

Kaye, J.D., J. Lowenstamm, J. Vergnaud (1990). Constituent structure and government in phonology. *Phonology* 7. 193-231.

Liberman, A. M., I. G. Mattingly (1985). The motor theory of speech perception revised. *Cognition* 21. 1-36.

Lindsey, G. A., J. Harris (1990). Phonetic interpretation in generative grammar. *UCL Working Papers in Linguistics* 2. 355-369.

McClleland, J.L., J.L. Elman (1986). Interactive processes in speech perception: The TRACE model, In D. E. Rumelhart, and J.L. McClelland (eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press. Vol 2.

Miller, G. A., P. E. Nicely (1955). An Analysis of Perceptual Confusions among some English Consonants. *JASA* 27. 338-352.

Mitchell, L., S. Singh (1974). Perceptual structure of sixteen prevocalic English consonants sententially embeded. *JASA* 55. 1355-1357.

Pols, L. C. W., L. J. Th. van der Kamp, R. Plomp (1969). Perceptual and physical space of vowel sounds. *JASA* 46. 456-467.

Rakerd, B. (1984). Vowels in consonantal context are perceived more linguistically than are isolated vowels: Evidence from an individual differences scaling study. *Perception & Psychophysics* 35 (2). 123-136.

Rakerd, B., R. R. Verbrugge (1985). Linguistic and acoustic correlates of the perceptual structure found in an individual differences scaling study of vowels. *JASA* 77. 96-301.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika* 27. 125-140, 219-246.

Shepard, R.(1972). Psychological representation of speech sounds; in David, Denes, *Human communication: a unified view*, pp. 67-113. New York: McGraw-Hill.

Singh, S. (1976). *Distinctive Features: Theory and Validation.* Baltimore, MD: University Park.

Singh, S. (1970). Interrelationship of English consonants. In B. Hala, M. Romportl, and P. Jannota (eds.), *Proceedings of the Sixth International Congress of Phonetic Sciences*. 825-828. Publishing House of the Czechoslavak Academy of Sciences, Prague.

Singh, S., Black, J. W. (1966). Study of Twenty-Six Intervocalic consonants as spoken and recognized by Four Language Groups, *JASA* 39. 372-387.

Singh, S., G. Woods (1971) Perceptual structure of 12 American English vowels. *JASA* 49. 1861-1866.

Singh, S., D. R. Woods, G. M. Becker (1972). Perceptual Structure of 22 Prevocalic English Consonants. *JASA* 52. 1698-1713.

Soli, S. D., P. Arabie (1979). Auditory versus phonetic accounts of observed confusions between consonant phonemes. *JASA* 66. 46-59.

Stevens, K.N., S.E. Blumstein (1981). The search for invariant acoustic correlates of phonetic features. In Peter D. E and J.L. Miller (eds.) *Perspectives on the study of speech.* Hillsdale, N.J: Lawrence Erlbaum Associates: 1-38.

Terbeek, D. (1977). A cross-language multidimensional scaling study of vowel perception. *UCLA Working Papers Phonet.* 37. 1-271 .

Wang, M. D., R. C. Bilger (1973). Consonant Confusions in Noise: A Study of Perceptual Features. *JASA* 54. 1248-1266.

Wickelgren, W. A. (1966). Distinctive features and Errors In short-term Memory for English consonants. *JASA* 39. 388-398.

Williams, G., W. Brockhaus. (1992). Automatic speech recognition: A principle-based approach. *SOAS Working Papers in Linguistics and Phonetics* 2. 371-401.

Wilson, K. V. (1963). Multidimensional Analysis of confusions of English consonants. *Am. J. Psychl.* 76. 89-95.

Wish, M. (1970). An INDSCAL analysis of Miller and Nicely consonant confusion data. Presented in 80th Meeting of the *Acous. Soc. of Am.* Nov. 3-6, Houston.

Wish, M., J. D. Carroll (1974). Applications of INDSCAL to Studies of Human Perception and Judgement. In E. C. Carterette and M. P. Friedman (eds), *Handbook of Perception*. New York: Academic. Vol. 25. 449-491.