Journal of Quantitative Linguistics
1999, Vol. 6, No. 2, pp. 117–148

0929-6174/99/0602-0117$15.00

Vector Phase Space for Speech Analysis via Dimensional Analysis*

H.M. Hubey
Montclair State University

ABSTRACT

A vector space using dimensional analysis is produced in which one can show all the phonemes/phones of all languages. Vowels, and consonants can all be shown in this phase space. Furthermore, the three-dimensional vector space for vowels, which in simplified form can be shown to be related to the distinctive features, can also be compressed to fit in this phase space for speech. This phase space can be shown to be both based on articulatory/geometric considerations, that is the two-tube model of Fant and Stevens, and also on the quality/perception arguments based on formant studies (Peterson & Barney, and Clark & Yallop). It can be used to clarify and unify many linguistic phenomena such as child language (Anderson, Jacobson), aphasia, sonority, the cardinal vowel diagram (Jones, Ladefoged), diphthong trajectories (Carre & Mrayati). It is shown that the sonority scale is directly correlated with this space in that sonority is related to the distance of the phones/phonemes from the origin. Hence sonority is a function of the magnitudes of the vectors (phonemes/phones) of this space. Diphthong and vowel confusion that crops up when using Artificial Neural Networks (Kohonen) for vowel recognition is easily explicable in this space. The fortition-lenition phenomena and phonological strengths (Foley) are nothing but vector phenomena in this space. The reasons that almost all languages have the phonemes /ptskn/ can be clearly shown in this space as splitting up the available phonological phase volume into nearly equidistant volumes. It is shown that this space is ideal for the discussion of such seemingly disparate phenomena as assimilation, metathesis, haplology, and dissimilation. In short this phase space is the natural phase space for speech since it

1) provides unification for as diverse phenomena as articulation, acoustics and perception of linguistics;
2) explicitly shows how to display spatio-temporal phenomena;
3) points the way to improvements using empirical measurements.

INTRODUCTION

Properties of Consonants and Vowels

Doing science consists of stages of: *analysis* (taking things apart and analyzing the parts) and *synthesis* (putting back the parts and understanding the operation of the whole in terms of its parts). In the standard analysis of the building blocks of human speech (phones and phonemes) whatever is necessary to understand system properties of language is not normally attributed to these parts, so that synthesizing a system (a whole) in which one can demonstrate the underlying unity of seemingly diverse phenomena is not achievable. However, a multidimensional metric space can be constructed in which one can demonstrate the essential unity of the articulatory, acoustic and perceptual perspectives of speech, as well as the standard divisions of speech sounds into vowels, consonants, and semivowels. Furthermore, one can display spatio-temporal phenomena such as sonority, aphasia, locus equations, phonotactics, haplology, metathesis, fortition/lenition, dipththong and vowel confusion. The analysis must, necessarily, take place at different scales, and since phenomena at different scales display relationships of different kinds, ideally we should be able to obtain macroscopic behavior equations via appropriate approximations from the microscopic equations. However, because of the difficulty of obtaining such equations, one must often resort to substitutes for explanations which are only intuitively understood. There is often a trade-off

*Address correspondence to: H.M. Hubey, Department of Computer Science, Montclair State University, Upper Montclair, NJ 07043. E-mail: hubey@mail.montclair.edu.

in precision and accuracy. We would like to be able to extend the concept of continuous orthogonal vector spaces to consonants or contoids. However, the formants for consonants do not exist, almost by definition. They will certainly not exist in the sense of the formants of vowels. In fact, this can be corroborated easily (see Edwards, 1992). Of course, we can use the distinctive features spaces (Hubey, 1994), however, the dimensionality is too high. We would like to be able to generate a broad transcription space to describe the consonantal sounds in ways similar to vowels. This practically limits our dimensions to two or three. One of the most obvious characteristics of consonants (in contrast to vowels) is that the articulatory organs move in time, whereas vowels are *steady-state* constructs. The concept of steady state comes from physics in which the concepts of equilibrium are not sufficient to capture the idea of the existence of a dynamic state in which one can identify certain constants or invariants of the motion (in an appropriate phase space). In the case of vowels, the production of the vowel is simplified by breaking it up into three stages; initialization, steady-state, and end. This is similar to the description of musical notes in which there is attack, steady-state, and decay. During the production of the vowel the shape of the power spectrum is important but not exactly where the power is concentrated (as long as it is in the lower bandwidth of the human hearing range) since the basic pitch of sound differs, with women and children producing relatively high pitches compared to men. We may think of the vowels as analogues of carrier frequencies used in electromagnetic communication. Indeed, even more apt is the analogy to the dial tone used by telephone companies which is DTMF (dual-tone multi-frequency) since the formants are peaks of the power spectrum so that they would correspond to the multi-frequencies of telephone dialing.

In contrast, the constants are modulations of the carrier signals and are by definition dynamic. The only invariants by which we can detect the consonants are the changes in the carrier signal which is seen in the power spectrum. In any case, the necessity of a dynamic description is obvious not only in the motion of the articulators needed to produce the consonants but also explicitly accepted in their descriptions in terms of such motions. This dynamic property of consonants is shared by the semivowels, diphthongs and glides. However, at a very broad level we can also imagine a class of consonants that share another property with the vowels. Certain consonant classes, in particular the nasals such as {/m/,/n/}, the liquids such as {/r/, /l/} (referred to as glides by some), the voiced fricatives such as {/v/,/z/} are ɪ-colored (the Turkish-ɪ is used for schwa-like phones/phonemes and it is writ-

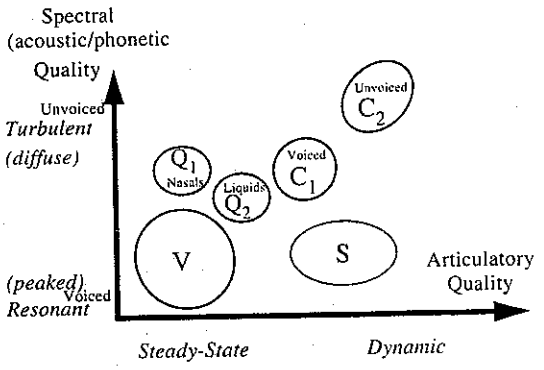

Fig. 1a. Four-fold Way: The division of speech sounds into basic classes. The two dimensions, the abscissa and ordinate are not really orthogonal.

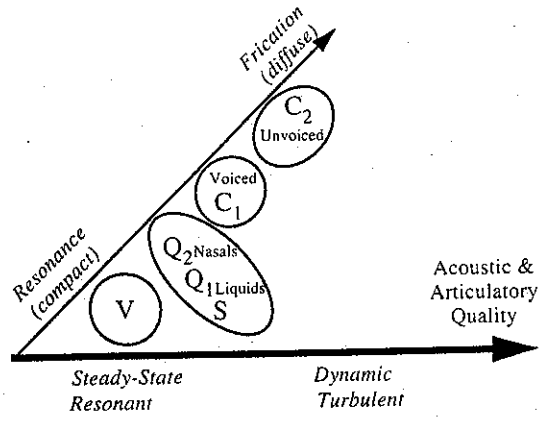

Fig. 1b. Nonorthogonal Spaces: The basic speech sound sets can be represented along a set of non-orthogonal axes in very broad transcription/categorization.

ten in bold to denote that it is a vector; see Appendix VI) and may be called *quasiconsonants* in the same sense as semivowels and diphthongs and have some dynamic properties such as consonants. When one deals with coarse-grained description of phenomena one expects less precision in the descriptions. At the highest (most coarse-grained) explanations of the underlying physical units of speech the primitives used are consonants (C) and vowels (V), with semi-vowels (S) playing the intermediate role. With the introduction of the set of quasi-consonants, a four-way division can be created based on fundamental properties of speech sounds. A suggestive space is given in Figures 1a and 1b.

Generally only the first three formants are used in speech synthesis and analysis because they are more than sufficient to approximate the qualities of the vowels. Since this requires at most a three-dimensional vector space, we can easily generate eight vowels as the corners of a parallelepiped in this space and approximate them using distinctive features (see Hubey, 1994, 1999)). The eight Turkish vowels which denote an almost perfect match for this space are used as vectors, (see Appendix IV; see also Hubey, 1996b). The property that the quasiconsonants share with the vowels is that they are also *steady-state* sounds in that the articulatory organs do not move in time, although the DOF (degrees of freedom) of quasiconsonants such as /l/,/z/ are zero whereas the nasals, /f/ and /v/ leave the tongue free to move about. That these consonants seem to have vowel like qualities can be seen in their power spectra (see for example, Edwards, 1992). There seem to be high-peaks at a low frequency with an exponentially decaying amplitude which is what we would expect from an ɪ-colored vowel or consonant; that is, the spectra resemble an ɪ with enough noise (low signal-to-noise ratio) to bury the signal, (see Appendix I for a better description). Of all the vowels the most neutral, and in a way the most *well-behaved* vowel as can be seen on the power spectrum is the ɪ (Edwards, 1992). Its formants' amplitudes drop off exponentially as if due to a frequency dependent attenuation. The quasiconsonants all seem to show some evidence of this phenomenon, as can be seen in Edwards (1992). In addition, the fricatives are also steady-state sounds, however their sound quality does not show any evidence of vowel like quality because of the high level of turbulent noise which drowns out the peaks that would have been produced by the specific configuration of the articulators or the shape of the vocal tract as an acoustic system. The plosive groups (especially the unvoiced) would best be modeled in the time-domain as Dirac delta functions. Of course, this implies that the power spectra would contain energy at all frequencies and thus would be flat. The voiced plosives are differentiated from the unvoiced essentially by the magnitude of the difference in time between the voicing and the plosion so that they would also be in the consonantal group. Thus from the basic division along the vocoidal-contoidal continuum we can derive, as a zeroth-order approximation, a symmetric four-way division; vowel (V), semivowel (S), consonant (C), and quasi-consonant (Q).

TOWARDS A SPACE

We know that one of the fundamental determiners of the quality of a speech-sound is the location of the primary constriction (Fant, 1990; Catford, 1988; Ladefoged, 1962; Lass, 1984). Recent research has pointed out that what we knew intuitively is correct, i.e., stop place changes recognition (Stevens, 1989). It thus seems that we already have two dimensions in which to represent the consonants. If we denote by S the size of the stricture (i.e., the size of the primary constriction) for the consonant, then one of our dimensions would be $Y=|\partial S/\partial t|$ where the vertical bars indicate the absolute value of the derivative. It will be shown later that it is not necessary for the derivative to be partial; indeed it might be more useful otherwise. We can use the location of S as another dimension, say X, essentially a mapping starting from the lips (for the bilabial consonants) and extending back towards the soft palate and pharynx. Although we consider this to be a single dimension extending in curvilinear fashion from the lips toward the pharynx (see Peterson & Barney, 1952) it will

be shown later that X should really have more (physical) dimensions.

The third dimension for the consonant 3-D vector space (Z dimension) would have something to do with *sound quality* or *airflow quality*. It is essentially the dimension that would distinguish the turbulent-chaotic quality of fricatives and sibilants from the more vowel-like quality of the quasi-consonants (see Appendix I). The particular boundary between these sounds is not clearly delineated since some phonemes such as /v/ and especially /j/ very clearly show evidence of both turbulent or chaotic flow (friction) and laminarity (resonation or periodicity). As is well known, the simplest kind of periodic motion is described by the harmonic oscillator differential equation whose solution is sinusoidal. The damped harmonic oscillator (the RLC circuit) is the primary equation used in phonetics to model the resonances which comprise the peaks in the power spectrum which are the formants of vowels (Fant, 1990). The digital versions of these resonators are used in speech synthesis programs such as MITalk (and whose source code was provided by Klatt, 1980). In increasing order of complexity the equations that describe these would need noise added to produce the fricatives. A short description of this process is in Appendix II.

The analogy with fluid flow is that at low velocities (actually small Reynolds' numbers) the flow is laminar in the sense that the fluid particles flow along sheets (layers). At values of Reynolds' number beyond about 2,000 the flow becomes turbulent. Turbulence is noisy in the sense that the deterministic regime is gone and the motion of the particles can only be described stochastically. But it is known that sounds (pressure waves) can scatter from turbulence and there are various methods of extracting signals scattered from turbulence. The example above (Fig. 2a) shows the relationships of some phonemes but is not drawn to scale. The third dimension (voicing) can be included in the discussion above, if we include the larynx as part of the geometric parameters or the articulatory organs. Therefore, we can use some kind of a weighted average of the rates of changes (i.e., time derivatives) of the articulatory apparatus (including the rate of change of the primary stricture) to put the voicing dimension along with Y.

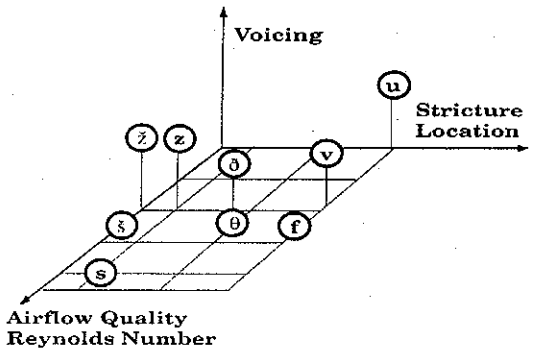

Fig. 2a. An Intuitive/Suggestive Space. At this level of approximation/simplification one can take the Reynolds' number used in fluid dynamics to define what is meant by 'airflow quality'. It delineates the boundary between laminar and turbulent flow of fluids. The circles may be thought of as the most representative of the set of objects which they define. In that sense the circles above are neither phonemes nor phones. See Appendix II for clarification and also for definitions of the symbols used.

CONSONANT VECTOR SPACE AND DIMENSIONAL ANALYSIS

Figure 2a shows some common consonants plotted in the X, Y, Z dimensions (i.e., essentially *Primary Place of Stricture, Rate of Change of the Articulators,* and *Quality of Airflow*). The drawing is not to scale. It is been distorted to give a general idea of the positions of some of the common consonants. It will be shown in the next section that this arrangement is not fortuitous but obeys very fundamental laws of physics. Real world phenomena take place in *space-time*. These are the *fundamental dimensions;* that is, it takes three space coordinates (dimensions) and one time dimension to describe mechanical events. However in dimensional analysis, the space dimensions are collapsed into one dimension, usually denoted by L. Thus the di-

mensions of area L^2 and L^3 describe three-dimensional space. The time dimension, of course, is denoted by T. However another level of abstraction is needed to describe the fundamental processes of physics. For mechanics we need one more; usually Force F, or Mass M. They are not independent since they are related by Newton's formula F=ma. For electromagnetic phenomena we need another called Charge Θ, and for thermodynamics, Temperature, Θ. Speech is not an electromagnetic or thermal phenomenon but a mechanical one and needs three dimensions F, L, and T. *Dimensional analysis* is often used in fluid dynamics (see for example White, 1979) where the processes are too difficult or complex to describe simply because they involve many variables and are highly nonlinear. Dimensional analysis helps in identifying groups of variables for which experimental relationships should be sought. The idea has proven highly successful in experimental fluid dynamics. Dimensional analysis was first proposed and used by Buckingham and the method used to find the dimensionless groups is called the *Buckingham Pi Theorem.* Speech sounds involve both wave mechanics and also turbulence effects (which are best modeled mathematically as noise, although the new mathematics of chaos could be useful for digital production of noise for fricatives). It is because of this complexity that dimensional analysis has been used in this paper to integrate the various aspects of speech sounds.

If we examine the dimensions of the coordinates X, Y, and Z, we will see that we are very close to what dimensional analysis would have yielded. The Z coordinate which we called airflow quality corresponds to the dimensionless group in fluid dynamics which discriminates essentially between laminar and turbulent flow. It

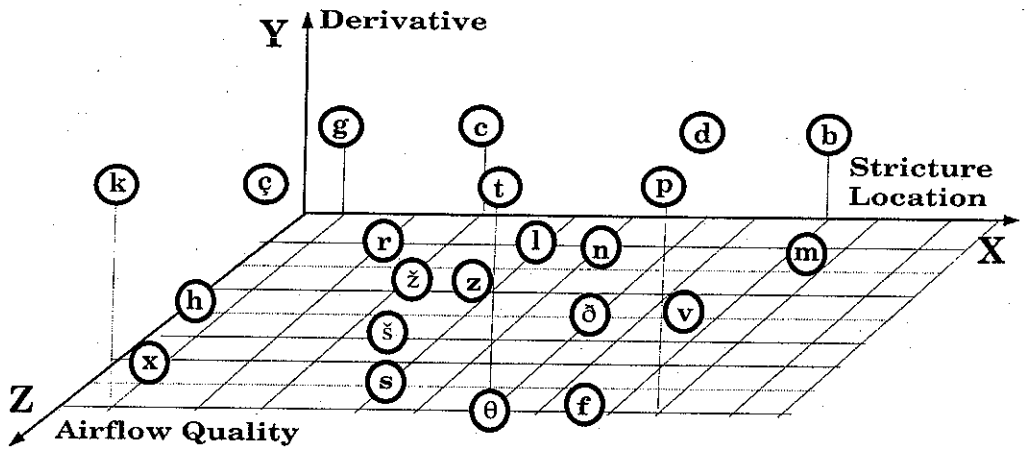

Fig. 2b. First attempt at a more complete three-dimensional space. Various expressions are used in phonetics and phonology to describe 'qualities' of speech sounds, and different groupings are recognized by different authors. Combinations of various phrases and words used in the literature are used to describe here in addition to the IPA alphabet (wherever possible). The Z axis phonetics would translate roughly to words such as *sibilant,* and *fricative* at the high end. At the lower end because of laminarity of the flow, the sounds would be called *resonant* (highly peaked power spectrum), or *periodic.* The sounds having to do with or caused by rapid changes in time of the articulators are usually described by *plosive, obstruent,* and *stop.* The word *continuent* is usually reserved for sounds which in this paper are referred to as *quasiconsonant* (or *steady-state* without regard to other qualities). The words *liquid* and *nasal* refer to the consonant midway between the two extremes and in which one can observe *resonance* and *friction* or *turbulence* (i.e., noise) in steady-state. In addition, *contoiod,* and *vocoid* are used to refer to consonantal and vowel like sounds. In addition, descriptions such as grave, acute, diffuse, compact, glide, sonorant, oral can be seen in the literature but are not used here except when required to refer to other author's writings.

is called Reynolds' number and is given by Kv/υ, where K [L] is the characteristic length; v [LT^{-1}], velocity and υ [L^2T^{-1}], kinematic viscosity. The dimensions of the variables are given inside the square brackets. No specific references were made to the kinematic viscosity because speech only takes place in air and not in other material. The Y coordinate is the time derivative of an area (stricture); thus its dimension is L^2T^{-1}. The third coordinate X is essentially the place of the stricture in one dimension. However, it was noted that it would have been better for it to have the dimension L^2. Thus we can take the height dimension into account by modifying the horizontal place coordinate by using the horizontal coordinate multiplicatively, in ways similar to Peterson and Barney (1952). All we have to do now is to multiply by a *characteristic frequency* ω (whose dimensions are not cycles/sec (Hz) but radians/sec [T^{-1}] so that X has the dimensions L^2T^{-1}. It is clear now why the correction via a function of ω (having the dimension T^{-1}) is necessary. It is common knowledge that in speech studies it has been found necessary to account for the differences in pitch of various speakers. It is not the absolute formant frequencies but the relationships of the formants to one another that is used to distinguish the vowels (Nearey, 1978; Liberman & Blumstein, 1988). The method used, the so called *vocal tract length normalization,* performs this function, since the smaller vocal tracts result in higher fundamental frequencies for the speaker. Although the size of the glottis, the mass of the vocal cords and speaker specific characteristics influence the fundamental frequency or pitch of the speaker, in general since the vocal tract acts as a filter, the resonance (frequency of greatest gain) is dependent on the length of the vocal tract. In simpler terms, the size of the tubes in the two-tube model of the voice production is a determiner of the fundamental frequency (aside from the characteristics of the medium). For a single tube such as an organ pipe the resonances occur at nv/2L, where L=length of the column, and v is the velocity of the acoustic wave (see for example Halliday & Resnick, 1978:441). Thus there is a functional relationship between the length of the vocal tract and the *characteristic frequency* (pitch) function proposed here. Hence the dimensions of X are L^2T^{-1}. For a better approximation, even if we tried to use dimensional analysis to find another proxy for L, we would still need to consider some characteristic length such as the size of the glottis, mass of the vocal cords (which is a function of volume) or a weighted average of all of these. In such cases we might need to use fractional dimensions, which is not done here. We can now use this knowledge to redefine the coordinates of the *consonant vector space* to include vowels and semivowels. It was noted that the Y coordinate need not be only a partial derivative. Indeed, the partial derivative will not be able to account fully for even the consonants and especially for certain consonants such as the voiceless affricate /tʃ/ or the Turkish palato-alveolar fricative /ç/ since the motion of the articulators (the tongue in this case) is more complicated than what the partial derivative indicates. We should write the stricture function S not as a function of time only, but as S=S (X (t),t). The time derivative then is:

$$dS/dt = \partial S/\partial t + (\partial S/\partial X)\,(dx/dt) \qquad (1)$$

Furthermore, we can extend the definition of Y by defining it as a weighted average of the sums of the total derivatives of the strictures involved in the articulation, not only the primary stricture. This means that we are now accounting for the change in the vocal cords in the derivatives, therefore the distinction in this phase space between the consonants and vowels can be made, as well as the voiced and unvoiced vowels. Since the average of a sinusoidal wave is zero, the average that is used for the motion of the vocal cords should probably be another type, say a root-mean square (rms). All we have to do now is to divide both the X and Y groups of variables by the kinematic viscosity υ (as in the Reynolds' number, see, for example, White (1979)) and we will have a three dimensional vector space consisting of three dimensionless quantities; *Reynolds' Number*, and the two dimensionless groups or numbers that we just derived. If we denote the horizontal length as λ and the vertical dimension as η; the strictures as S_1

and S_2 (using the two-tube model), then we have;

$$X = \lambda\eta\omega/\upsilon\ [L^2T^{-1}]/\ [L^2T^{-1}] \qquad (2)$$

which is dimensionless. Y coordinate for two strictures could then be

$$Y = [d/dt\ \{|S_1\ (\lambda,\eta]| + |S_2\ (\lambda,\eta)|\}]\ /\upsilon \qquad (3)$$

which is also dimensionless. We might also include the third stricture (i.e., the glottis) explicitly in the determination of Y. Since all three coordinates are divided by the kinematic viscosity, in practice it might be eliminated so that the coordinates will be given by the (effectively) dimensionless numbers:

$$X = \lambda\eta\omega;\ Y = [|dS_1/dt|+|dS_2/dt|]_{rms};\ \text{and}\ Z = K\nu \qquad (4)$$

It should be noted here that the K (in Z), the characteristic length, is also a good candidate for some kind of vocal tract normalization or it may be used as a characteristic of stricture place.

These numbers are only suggestive and require improvements based on actual experimental measurements from many languages in the same way that dimensional analysis suggests the natural groupings of variables which should be used by the experimental fluid dynamicist. For example, for the bilabial plosives the rate of change of the constriction can be positive, negative or both. Thus, if we were to use only the absolute value, we would have a reasonable approximation. On the other hand, if the acceleration is not constant, we would need to use an average.

Since the opening and closing are of different signs the average would be zero unless we used a root-mean-square kind of an average. As for some of the specifics on friction and more easily measurable physical parameters such as pressure, Stevens, along with others, has done research on relationship of friction to pressure drop across a constriction (which can be found in Lieberman (1988)).

It is difficult to represent some of the differences of the phonemes due to friction, in a three-dimensional space since we really need a much higher dimensional space (see conclusion and

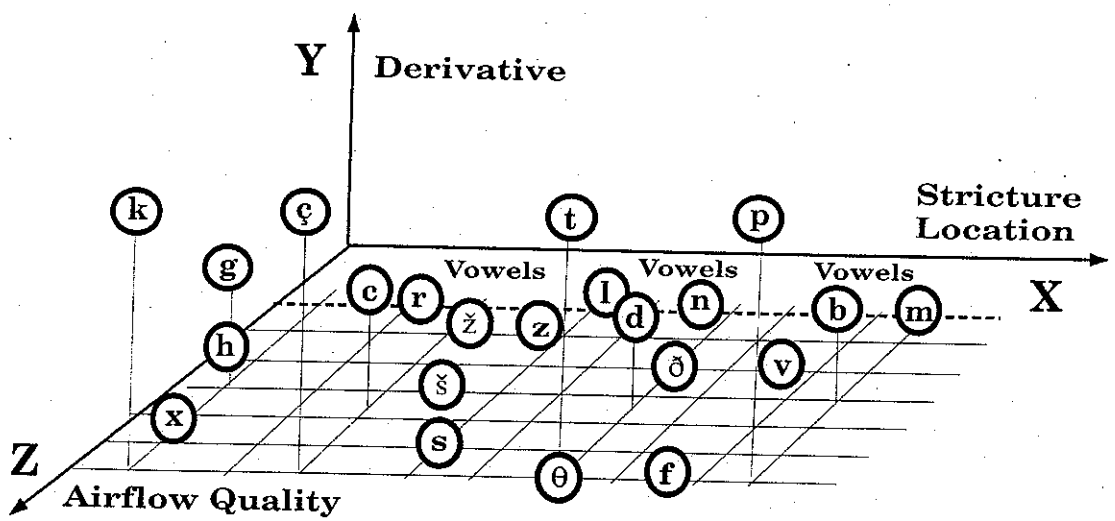

Fig. 3. Slightly better placement of the consonants in phase space. The vowels are steady-state and belong on the ZX plane. Furthermore, they are highly resonant and therefore cluster around Z=0, so that the voiced plosives must be moved as shown. The phonemes /ptskn/ essentially define the boundaries of the volume that the consonant occupy except that some voiced plosives seem to be missing to completely define this volume. These are also reasonably evenly spaced in this space.

discussion). Some kind of a weighted average conceptual scheme must be used. The placements shown in these figures are suggestive, and approximates since have been culled from many different areas of linguistics, as will be demonstrated over the next few sections. Chomsky and Halle (1968) mention that /ptskn/ are rarely absent in any language. It is clear now why even at the level of accuracy of Figure 3 it must be so. These five contoidal phonemes essentially define the boundaries of the volume of the consonant phase space, (as will be even clearer in Fig. 5) and they roughly divide the phase space into equal intervals/volumes which has implications for distinguishability since the relationship between the articulator positions (and manners of articulation) and the quality of sound (i.e., their perception or acoustic characteristics) must be deterministic, although highly nonlinear. The dotted lines denote the volume in which the vocoids, the vowels, semivowels and polypthongs fall. The little circles are meant to be representative manifestations or instantiations of the sets that comprise the phonemes. Hence, neither the phoneme nor the phone symbols are used. In truth they are neither since the results must be generalizable to all languages so that they are representatives of some sounds which can be recognized to belong to natural clusters, and which may be split up differently in different languages, although we are using (American) English as a vehicle for explication of the ideas. It should be understood that the symbols really denote volumes in this space and that their boundaries are to be considered to be fuzzy, as in fuzzy sets (Hubey, 1999) or fuzzy vectors.

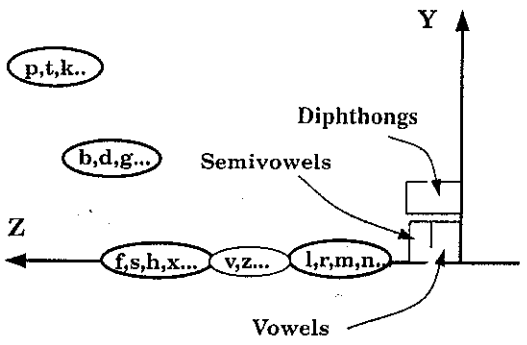

Fig. 4. Some possible groupings. The original coarse-grained categories in Fig. 1 can now be seen.

NATURAL GROUPINGS

The groupings in Figure 4 are two dimensional versions (orthogonal projections) of Figure 3, where the phonemes/phones close to one another have been grouped according to their characteristics. The figures give an indication of the way the phonemes/phones cluster in the *speech phase space*. They show very clearly that these divisions are those that have been described in various ways by phoneticians and linguists for centuries. It is not clear yet where /j/, /ʃ/ and /x/ really belong. In fact, judging from sound quality the /s/ does not seem to fit into its group either, i.e., with /f/ and /θ/. The final arbiter of the placements of the phones/phonemes has to be the results from acoustic measurements, for example, if the major frequency peak is lowered from about 5 KHz to around 2.5 KHz, the listeners' perception shifts from an /š/ to a /ʃ/. More about the fricatives can be found in Lieberman (1988, p.227). The grouping in the figure above is intended to show some natural clustering in the *phase space*.

The last phase space (Fig. 3) was derived from the original discussion on consonant spaces. However, after the discussion on dimensional analysis, the original dimensions or parameters were altered to take into account the various changes in Y, namely that it is the sum of the derivatives of the strictures and that the Z-axis has to do with the Reynolds' number. Now, the semivowels can be considered to be appended to the end (or beginning) of vowels with which they form diphthongs, thus there is a motion of the articulators so that their Y-values are not zero. The vowels are steady-state, therefore they should be solely on the XZ-plane extending very close to and partially mixed with the quasiconsonants since they also display some vowel characteristics (such as being /ɹ/-colored, and being steady-state (continuant)). The diphthongs are defined on the ZX-plane in the same area as vowels, except that their Y-values are not zero,

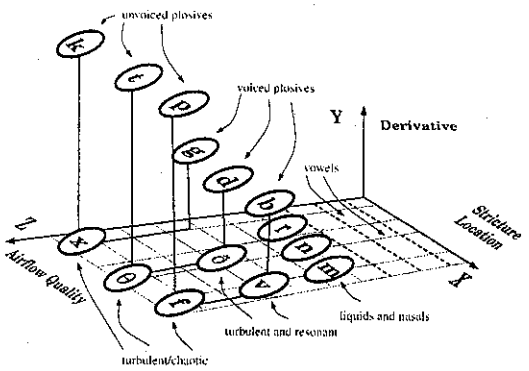

Fig. 5. Yet another version of the speech phase space. The vowels are close to the origin along X. The Y-axis is probably not to the same scale as the other axes. Thus, the voiceless plosives might be much higher than the voiced plosives.

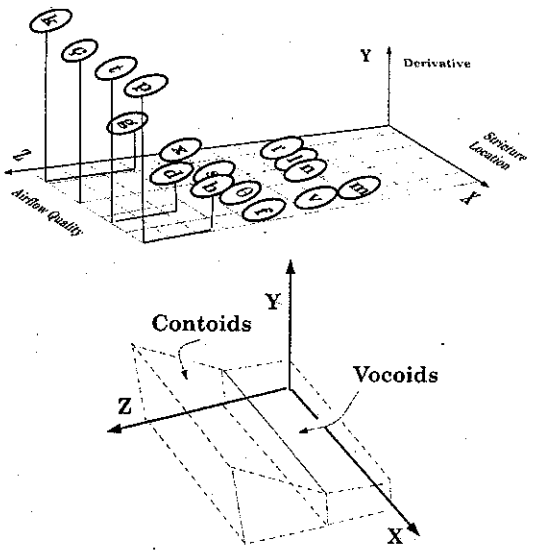

Fig. 6. Yet another figure suggestive of the speech phase space. The relative locations of the fricatives and plosives have been rearranged slightly. Vowels are not shown but are still along the X-axis. The only differences from Fig. 5 are essentially along the Z-axis since it is here that we note the differences in the friction and resonance of the various sounds. However, the general shape of the phase space can be seen already.

thus they will be located above the vowel range. The broken-line boxes indicate the vocalic sounds (vowels, semivowels, glides, diphthongs, and triphthongs). Finally, the voiced plosives should now be moved from the Z=0 range since we know that they contain both voicing (vowel-like sounds) and turbulence (the high frequencies that exist in short duration spikes, which can be modeled as Dirac delta functions in time). They should be moved somewhere between the vocalic and the consonantal sounds. All the changes are shown in Figure 5 from another view. It is even clearer now why /ptksn/ are rarely absent in languages. The /p/ is the extreme X (except for /w/); /k/ is the practical extreme for X (minimum) and /n/ defines the minimum Z. Any smaller value in the Z direction than /n/ would fall in the vocalic group. In the same chapter, Chomsky and Halle also remark that the /ɪ/ *(full schwa)* should be marked and should get a complexity of 2 along with the compounds like the /æ/. It would seem that the /ɪ/ is the most ubiquitous vowel especially in consonant-cluster laden languages like Slavic and to an extent English and other Indo-European, languages and it is spread through the region of much of the XZ-plane (except the vocalic parts which are covered by the specific vowels) since the quasi-consonants for all practical purposes are /ɪ/-colored. It would seem natural to have the /ɪ/ the least marked and the most natural to have in any system. It would seem that, from the place of pride that the basis vowels occupy, their position should be next, right after /ɪ/, and the others /e/, /o/ and /ü/ which can be constructed from the basis vowels could be next. Finally we are left with /ö/, since it requires all three basis vowels /i/, /a/ and /u/. (see Hubey, 1994, 1996, 1996b, and also Appendix VI)

Since these figures are not scaled, and indeed it is not possible to know with any degree of precision where some of these phonemes should go, some alternatives are given in these pages. It should be noted that, in general, the relative positions of the phonemes do not change appreciably, however it is not possible without more evidence to choose among the several competing alternatives. The liquids and nasals should probably be separated by a wider distance because of the more vowel-like quality of the nasals (i.e., nasal murmur).

LENITION, FORTITION, AND SONORITY

A concept that we will need for this subsection is that of a *vector* (see Appendix IV). An interesting usage of the concept of vectors will applied to lenition as can be seen in Lass (Lass, 84, p. 177) that gives the phonological rules for lenition and fortition as:

(a) Stop > Fricative > Approximant > Zero
(b) Voiceless > Voiced

Essentially the same results can be found in Foley (1977). These results can easily be shown to be derivable quite clearly and unambiguously in the phase space and are related to sonority. We only need two dimensions (although three would be better) and the concept of a Cartesian vector to show the essential results. The space shown in Figure 7 is a 2-dimensional subspace of the 3-dimensional phase spaces developed earlier. Indeed, the three-dimensional phase space can be considered to be a subspace of the many different feature-bundle spaces discussed in the literature with the caveat that these spaces are not orthogonal and the mapping might not be one-to-one or linear.

We can see immediately from Lass's hierarchy that (a) refers to a vector that points in the negative Y direction (Stop > fricative) which is $C_2 > Q_2$ or $C_1 > Q_2$. The second part of (a) refers to a vector that points in the negative Z direction (i.e., Fricative > Approximant). The third part of (a) is also a vector that points in the negative Z direction (i.e., toward the origin of the YZ-axes). Part (b) refers to a vector that points from C_2 to C_1 (Voiceless > Voiced) and thus is a vector that points in both the negative Y and Z directions. The vectors that show these concepts are shown in Figure 7. Since no measurements have been taken to indicate the scale of the phase space, and no mathematical definitions have been given, at best we can use the data from Lass (1984) and Foley (1977) as guides to make the phase space reflect reality as closely as possible. In the next subsections, data from child language development, aphasia and formant measurements will be used to fill in some of the gaps of the phase space. Meanwhile, it can be seen from the vectors above that all of these phenomena are easily describable in terms of the vectors representing the transitions. Thus, lenition is a vector pointing toward the origin. The sizes and shapes in Figure 7 are not important due to lack of scaling which itself is due to the lack of necessary measurements.

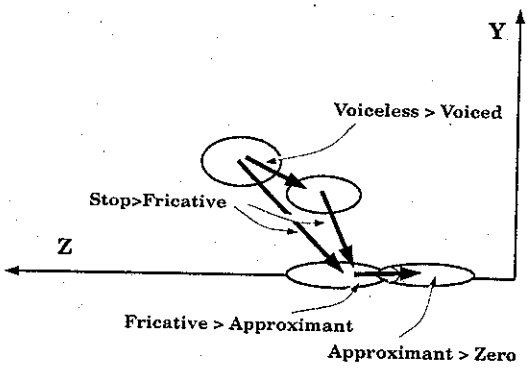

Fig. 7. Fortition, lenition and sonority. These concepts of weakening or strengthening of sounds are very fundamental in phonology (see for example Foley, 1977). The basic ideas can easily be 'explained' in terms of the vector phase space of this paper.

CHILD LANGUAGE DEVELOPMENT AND APHASIA

The study of child language development, although started by Jakobson in 1941, has amassed much data over time. It is summarized in Anderson (1985, p.131) as:

> ... all children begin with a minimal opposition of a single vowel (roughly [a]) and a single consonant (generally labial [p]). Consonantal distinctions arise with a difference between nasal ([m]) and oral ([p]) segment type; and subsequently with a split in point of articulation between grave (labial) and acute (dental) sounds. Within vowels, the first split is between compact (low) and diffuse (high) segments. With regard to manner of articulation, stops arise before fricatives, and both before affricates. The consonant/vowel dis-

tinction precedes the emergence of liquids or glides, and sonorant liquids precede obstruent liquids. Some distinctions, where they are to appear, arise only very late: e.g. nasal vs oral vowels; opposition between liquids; clicks, ejectives, implosives and other nonpulmonary airstream mechanisms, etc. The uniformity of the sequence in which these segmental distinctions are acquired seems quite general.

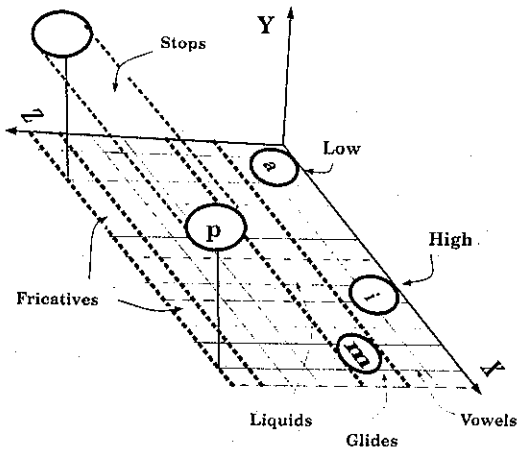

Fig. 8. Language development: Kindersprache and aphasia. Only some of the (early) speech sounds have been shown. The vowels are treated in the next section and also in the appendices.

We can attempt to sketch out where the vowels should fall rather easily from this description based on a very simple algorithm. All we have to do is to assume that children start by distinguishing the most different phones (i.e., those most distant from one another) and then continue to divide this volume into smaller pieces as their power of discrimination increases and as they listen to speech.

The sequence is roughly sketched out in Figure 8. Only the voiceless plosives are shown for the stops. It can be seen that the [a] and [p] start off with the maximal distance at the two extreme ends of the space. Then a nasal [m] is later introduced, then an [i] and as the process continues, it seems to further subdivide the phonological volume as if cutting a piece of cake into smaller and smaller pieces. We should note here that there could possibly be other reasons for the order of learning. The [a] is an open vowel, i.e., the fact that it is produced with the mouth open means that another channel of communication is available to the child, that is vision. Similar comments can be made about [b], [p], and [m]. The motion of the articulators for the back stops cannot be seen and the child needs more feedback before they are learned. Similar considerations might apply for the learning of the other vowels. Considered in this light, it is not surprising that a child's first utterance seems to be something like [ma], [pa], or [ba]. If the most important factor were intelligibility we would expect the '*supervowel*' [i] to be learned first. It has been shown by Nearey (1978) and Lieberman and Blumstein (1988) that the vowel [i] serves or can serve as means for the brain/mind to normalize/scale the other vowels with respect to pitch. It should be noted that the liquids and nasals fall in the same general region of distance from the X-axis and the figure does not mean to imply that [m] is a liquid. However, various reasons have been given at different times to justify grouping the liquids and nasals together (for example, Jakobson, 1990).

Aphasia seems to go in the reverse direction with the last learned being the first lost, thus it seems that there is a stack-like structure (i.e., Last In, First Out) in memory (Anderson, 1985). The same phenomena can be observed in boxers during matches. The first thing to go, after a hard punch, seems to be the last things learned (i.e., bobbing, moving from side to side, holding the hands up and finally the inability to stand up). Near death experiences where people see tunnels and bright lights could be due to similar brain processes. Furthermore, victims of aphasia never seem to make two featural mistakes (i.e., change two features at once) but rather only one at a time (Lieberman & Blumstein, 1988) which seems to lend credence to the usefulness of the binarity idea of distinctive features or from a different point of view the ability to come *close* to the phoneme. Aphasia is a complex process and its effects (whether it is Wernicke's or Broca's, see for example Lieberman & Blum-

stein (1988)) seem to hinge on the way the brain's memory and neural computation work. Thus not much more can be said about what aphasia implies for problems in speech production. A hint of the whereabouts of the vowels in the phase space has already been given in this section.

VOWELS IN PHASE SPACE

In the previous subsections, the vowels were left out of the phase space because of the difficulty of ascertaining their locations. It is difficult, for example, to decide a priori whether an /o/ should be near the front because of the rounding or near the back because of the position of the tongue. However, since the phase space has not only articulatory content but also an acoustic one, it is possible to draw inferences from several results, cull the results and put the vowels in the phase space. Some evidence comes from sonority and yet others from formants. For example, it is shown in another section that the sonority of the consonantal sound groups/classes/sets can be described essentially in a two-dimensional space.

We can extend the concept of sonority to the full three dimensions of the speech phase space. For example, Foley (1977) pursues an essentially sonority based course in his descriptions of *phonological strengths*. In his descriptions of the vowels, he derives the phonological strengths (Foley, 1977, p.47) of vowels as {i,1}, {e,2}, {u,3}, {o,4} and {a,5}. It is also quite interesting that the distances between some of the vowels (Foley, 1977, p. 78) can be derived directly from the binary three-dimensional representation of the vowels (see Appendix VI); for example, Foley gives the differences of phonological strength as |**a-o**| = 1, |**e-o**| = 2 and |**i-o**| = 3. In this connection, it should be mentioned that Gilbers (1992) in his network representation of segments not only uses a binary representation of vowels, but uses binary operations to derive vowels from others via operations of rounding, tensing, laxing etc. Gilbers (p. 130) also reaches the conclusion that 'we predict that unarticulated voicing, the articulatory correlate of schwa, is universal. In the area of first language acquisition, we consider schwa to be the first acquired vowel.' In a system of *markedness penalties* or *taxes;* he assigns zero penalty to the schwa (/ɪ/), small penalties to **i, u, a** and the largest penalty to **ö** (Gilbers, 1992:133) which is fully consistent with the results of this paper.

Table 1.

	i	I	ɛ	æ	a		U	u
$f_2 e^{-f_1}$	0	0.26	0.56	0.85	1	0.65	0.37	0.65
$f_2 e^{-f_1}$	1	0.79	0.68	0.6	0.17	0	0.12	0.02
$f_2 (1 - e)^{-f_1}$	1	0.58	0.30	0.09	0	0	0.076	0.001
$f_2 e^{-f_1}$	1	0.61	0.39	0.26	0.06	0	0.08	0.01
$e \times p^{f_2 - f_1}$	2.7	1.7	1.13	0.78	0.44	0.65	0.78	0.53
$f_2 \times (e - e)^{-f_1}$	1.7	1.12	0.66	0.23	0	0	0.15	0.02

Table 2

$f_2 (1 - f_1)$	i	I	ɛ	æ	U	u	(ɑɔ)	
$f_2 e^{-f_1}$	i	I	ɛ	æ	U	a	u	ɔ
$e^{f_2} e^{-f_1}$	i	I	ɛ	(æU)	ɔ	u	a	
$f_2 (e - e^{f_1})$	i	I	ɛ	æ	U	u	(a ɔ)	

The most important results that are necessary for placing the vowels in the phase space come from acoustic studies. For example, Nearey noted that the *front* or *acute* vowels ([i,I,e,æ]) have a high F_2 and the *back* or *grave* vowels ([a,U,u]) have a low F_2. The [i] and [u] with a high tongue position have low F_1 and [a] with a low tongue position has a high F_1 (Lieberman & Blumstein, 1988, p. 222). This implies that the function that we need to derive the placement of the vowels along the X-axis should decrease with increasing F_1 and increase with F_2 yielding some kind of a scaling along a *front-high* dimension. There are many ways of constructing such functions. Only a few simple forms will be given here to produce a rough *complete* phase space. The first step is in scaling the formants (as in Appendix VI)

$$f = (F - F_{min})/ (F_{max} - F_{min}). \qquad (5)$$

It is immediately clear that this scaling produces values which are dimensionally homogeneous regardless of the dimensions of the formants. This simple linear scaling was selected to show the possibilities of this type of normalization for producing a useful ordering for the vowels (also see Appendix VI). Table 1 shows the results of computations of some candidate functions.

Table 2 shows the ordering of the vowels according to the various computed values. In all the cases the various allophones of the acute vowels are near the front and the back allophones have low scores. The ordering essentially seems to show the distance from the origin (which is where the **a-o** apparently belong, radially outwards. The vowels which have approximately equal values are put in circles. Since

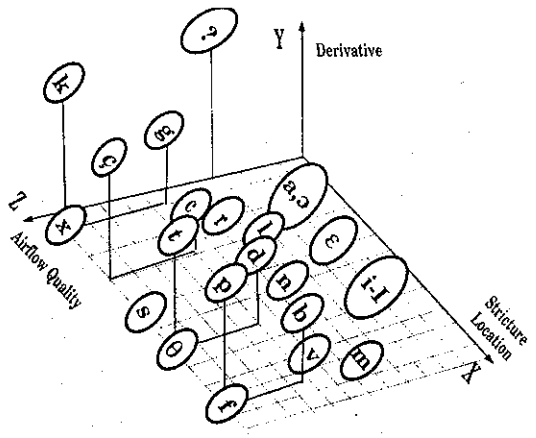

Fig. 9. The complete speech phase space. This is only one possibility showing a particular placement of the vowels. The fricatives might need to be re-arranged. The placement of the glottal stop is only suggestive.

these vowels do not fall on the corners of the perfect cube but rather are scattered about the corners of a distorted cube, this kind of accuracy (or lack of it) is expected. Thus we can produce a full phase space indicating the whereabouts of the vocalic sounds and phonemes of various languages. A rough sketch of the phase space with consonants and vowels is shown in Figure 9. Anything more accurate than this requires special experiments to determine the values of the phones/phonemes in dimensionless groups. Once again, it should be remembered that the drawing is not to scale since imputing distances using a linear Cartesian distance for sonority will cause problems. However, a three dimensional sonority scale can still be derived from this by weighting the coordinates.

DISTANCE, BIRTH OF NEW PHONEMES AND EXPERIMENTAL EVIDENCE FROM DIPHTHONGS

Ancient Sanskrit only had three vocalic phonemes /i/,/u/ and /a/ and eventually obtained an /e:/ and and /o:/ at a later time. Since it already apparently had the diphthongs /ai/ (or /ay/) and /au/ (or /aw/) we might wonder if there is a rela-

tionship that can be shown using the concept of distance. Figure 10 is suggestive of the first three formants of the diphthong /ai/ and the transitions from /a/ to /i/. The horizontal lines indicate the formants. The small arrows indicate a separation of 1 KHz. The formant data are from Peterson and Barney (1952). Since the beginning of the diphthong begins with a steady state /a/ (which means it is a vector in the n-dimensional formant vector space) and then ends up as another steady-state vector resembling an /i/ (or /y/), the whole phoneme must thus be represented as a vector transition (a dynamic vector or vector velocity) which implies that (1) vector derivatives and vector calculus becomes necessary and (2) it must necessarily pass through points of which some might belong to the volume of another phoneme.

The figure indicates what would happen if we substituted the formants for /e/ in the transition zone. The formants of /e/ fall in the zone where the transitions occur. It does not seem to be accidental. If we do the same thing for the /au/ diphthong we get a similar result, as can be seen in the figure above. Of course, the historical interpretation does not seem so clear cut. It is not clear if these changes were innovations or if new language speakers who did not have these diphthongs interpreted the changes as their own vowels /e/ and /o/. These ideas are discussed much more fully in Chapter VIII of Hubey (1994, 1999). Independent confirmation of the perception problems of diphthong transitions comes from an unexpected source. It is from speech recognition research using neural networks. Kohonen, who is a pioneer in research in the use of artificial neural networks for phoneme recognition, reports in his work in Japanese and Finnish that he has found that the network recognizes the diphthong /au/ in words like /hauki/ (meaning pike) as the /aou/ sequence and has found it necessary to introduce a phonological rule to derive /au/ from /aou/ (Aleksander, 1989, p. 35).

We may surmise from this that the /ö/ might have developed from the /üe/ diphthong the same way that /e/ developed from /ai/ and /o/ from /au/. This would imply that the ordinal vowels /e/, /o/ and /ü/ are rather closer to the diphthongs (i.e., the transitions). The experimental evidence from Carré and Mrayati (1991) shows the trajectories of the various diphthongs. Of course among these are the diphthongs /ai/ and /au/. Figure 11 indicates the paths that these diphthongs take in the two dimensional formant space. The space is not normalized but it is especially clear that the /au/ passes very near /o/. The figures above were indicative that this result was to be expected.

IMPLICATIONS FOR FORMANT-VOWEL SPACE

Figure 12 shows the rough ideal placement of the vowels from Hubey (1994) using the data from Peterson and Barney (1952) and Clark and Yallop (1990) for Australian English vowels, using the normalization given in Eq. (5) which can also be seen in Hubey (1996b). It is clear that this three-dimensional view of the vowels is

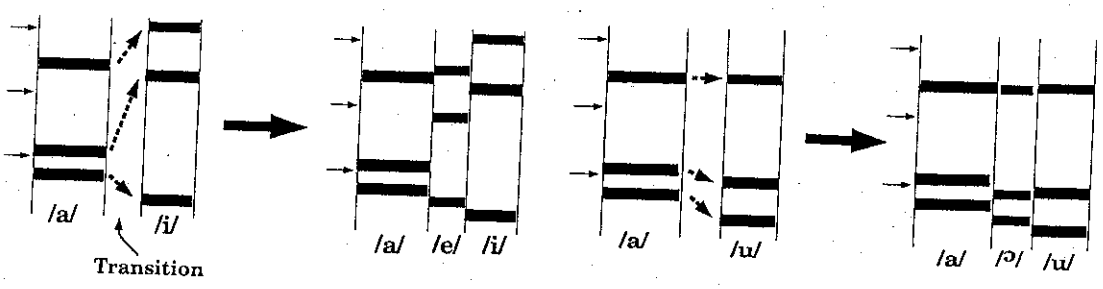

Fig. 10. Vowel formant transitions. Figures suggestive of the fact that diphtongs /ai/ and /au/ pass through or very near the vowels /e/ and /o/ respectively.

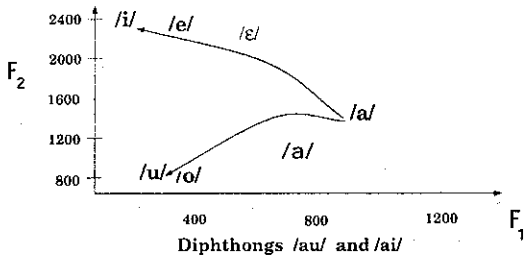

Fig. 11. Vowel transitions; after Carre and Mrayati (1991). One can see that the dipthtong /au/ passes near /o/ and that the dipthtong /ai/ passes near /e/. The results are also displayed graphically in Figure 10.

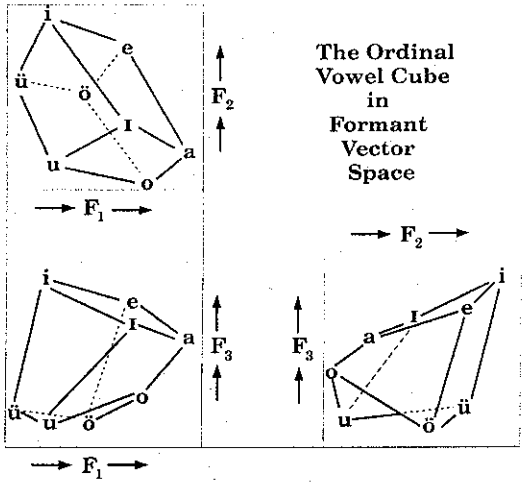

Fig. 12. The ordinal vowel cube (cuboid): This is a natural product of the empirically determinable acoustic characteristics of vowels of the American English language. From phonetic similarity we can extend the ideas to other languages.

an economic description of many linguistic phenomena. It fits in reasonably well with the traditional introspective vowel descriptions, the newer results from Ladefoged modifying the traditional cardinal vowel diagram, and the latest results of the experimental formant studies. Since it seems possible to represent not only diphthongs but also at least some of the consonants, such as the stops from the formant transitions, the importance of the formant vector space increases. It might be possible to represent speech sounds, with the addition of some other factors such as aspiration noise, burst amplitude, signal-to-noise ratio, within the formant vector space. It might also be possible to use both the formant vector space and the phase space together. The perceptual distance between the diphthongs and the ordinal vowels seems to imply that the ordinal-cube is not really cubic but that it is distorted (and also rotated) in the formant space. This is consistent with the results of the previous sections. We can try to construct this ordinal-cube in the normalized formant (vector) space by using all the information now available. The shape of the ordinal vowels in the formant space is shown in Figure 12 (a fuller description of its derivation can be found in Hubey (1994, 1996b). The side view (i.e., F_2 vs. F_3) shows some discrepancy with the formant data of Peterson & Barney and Clark & Yallop. The [e] has been slightly displaced to show more clearly the shape of the cubic structure. The [ü] does not show up in the formant studies and its position was estimated from various hints as alluded to in this chapter. It is hoped that a better normalization algorithm or samples of phones more representative of the eight ordinal phones/vowels, say from Turkish, will yield a better fit. As can be seen, the near-cubic shape of the eight vowels resembles the modified vowel diagram given in chapter I of Hubey (1994). A much simpified version using bitstrings and distinctive features which is related to this 'ordinal vowel diagram' and also to the linear normed vector spaces for vowels based on the first three formants, can be seen in Appendix VI.

The main problems in constructing this figure is that the [a] and the [o] do not separate well and that there is no [ü] or [I]. In addition, the [i] in English *(heed)* is diphthonged as is the [u] *(who'd)*. Moreover, the sample for the [o] is really from the open-o (as in *hawed*).

SONORITY AND SCALING

The idea of a sonority scale can be explained, first, directly from the graph since the sonority scale seems to go from the vowels toward the

plosives. So the scale is essentially the distance from the origin of the axis, the voiceless plosives being the least sonorant and the low vowels being the most sonorant, thus being inversely proportional to the distance from the origin (at least in the two dimensions as shown). The physical explanation, of course, is that in wave propagation, the high frequencies dissipate faster and thus low frequencies go further. The voiceless plosives are highly spiked and thus contain high frequencies. For example, an ideal spike of zero duration is a Dirac delta function and its Fourier Transform is constant implying equal power at all frequencies. The sonority then can be expressed simply as either one of:

(6a)	$\sigma = 1 - R$	$0 \le Y,Z \le \sqrt{2}/2$
(6b)	$\sigma = 2/\{1+R\} - 1$	$0 \le Y,Z \le \sqrt{2}/2$
(6c)	$\sigma = e^{-R}$;	$0 \le Y,Z \le \infty$
(6d)	$\sigma = \log (R)$	$1 \le Y,Z \le \infty$
(6e)	$\sigma = (\alpha Y^{2n}+\beta Z^{2n})^{1/2n}$	where $\alpha+\beta=1$ and $R= (Y^2 + Z^2)^{1/2}$.

Obviously, these functions might have various coefficients depending on the units and dimensions used. It should be noted that the definition is only two-dimensional but extension to three dimensions is straight forward. It would thus seem, recalling the positioning of the glottal stop that it might belong closer to the plosives than toward the origin which makes it closer to vowels. It would also be possible to represent the vowel transitions as vectors in a three-dimensional space. Since the pure ordinal vowels can be represented as the corners of three-dimensional cubic structure in the 3-D formant space, the transitions between two vowels, whether it is a diphthong or a consonant should still be recognizable as a velocity (i.e., transition). Thus, if only formants are used (extracted from the signal in short intervals, say about 10 ms intervals), then the transitions will be represented as vector derivatives in the formant space. Some consonants show changes in the higher formants although some clearly show transitions in the lower formants. In the phase space of the previous sections, it is clear that vowels are statically representable. It is also possible to view the formant transitions as induced by the stop consonants as studied by Stevens and Blumstein (1978), Lieberman (1984), Lieberman and Blumstein (1988) and Blumstein and Stevens (1980). It has been known since the earliest studies at Haskins Laboratories (Lieberman, 1984) that the stop consonants cannot be isolated in speech from the vowels and that they show up as transitions of the formants in CV syllables such as [ba], [bu], [bi], [da], [du], [di], [ga], [gu], [gi]. Much research has been conducted to look for acoustic invariants representing the consonants in these syllables; and it is still continuing. The placements of the various phonemes in the phase space is consistent with the sonority scales or acoustic power, given in the literature, for example, from Fry (1979) or Levitt (1978) and which can also be found in Edwards (1992).

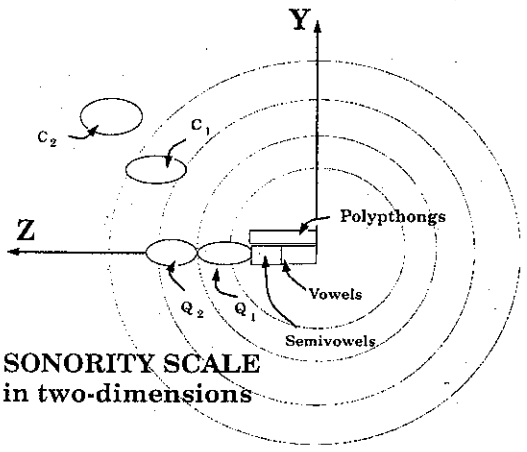

Fig. 13. Suggestive distances from the origin and sonority: The strength of vowels on the sonority scale can be seen to be inversely proportional to the distance from the origin as can be seen in the phase space.

CONCLUSION AND DISCUSSION

There is something about the phase space that strikes people as odd; is it articulatory or acoustic? The simple and correct answer is that it is both. There is no reason to be surprised about why the articulatory and acoustic parameters

should map to one another. If, in fact, it were not so, then we would be really surprised since the particular acoustic manifestation of a sound is dependent on the articulation of it. The likelihood that this mapping would be highly nonlinear is taken for granted since it is common knowledge that different articulations can give rise to the same spectral pattern or acoustic/phonetic output (Ladefoged, 1962, 1971, 1990) And the dimensionless numbers (which are the dimensions of this space) are really this nonlinear mapping! The nonlinearity is absorbed into the dimensions of the 3-D space and the result is a more tractable space; a simple one, yes. The proposed phase space has both acoustic and articulatory content and why should it not? If there were no degree of correlation between articulation and acoustics, how then can we produce the sounds we want? If there were no degree of correlation between acoustics and perception how then can we have any regularity and use speech as a communication tool? Speech phenomena certainly possess extreme complexity and it is exactly because this complexity is of both fluid motion (laminar and turbulent) and wave patterns imposed upon it that dimensional analysis is required.

Of course, the simple space has limitations; it cannot represent geminates, ejectives, trills, clicks, and pharyngealization because it is not complex enough. To do that we would need more dimensions. It might require several tens (or thousands) of nonlinear stochastic differential equations to reproduce even some of the complexity inherent in speech production. The strength of the model is that it has something that others do not; none of the papers in any phonetics journals as yet shows any rhyme or reason for the data being collected or the scatter plots that are being produced. The most fundamental principle of physics is that the terms of any equation must be dimensionally homogeneous. Dimensional analysis will produce the correct result. This phase space cannot give high resolution descriptions because it is of low dimensionality.

Some of the simplifications are that specifically; what is being done not only for the exotic sounds as above but also even for some of the more common ones (such as [ʃ], (j) and the nasals) is that they are being forced into the 3-D space by tinkering with the extra dimensions that would be necessary for representing them properly. For example, Fant (1990) included the information in the third formant by modifying the second formant values to be able to use only the first two formants. Similarly, the extra dimensions of the vowels have been squeezed into a single dimension by ignoring the third formant and collapsing the first two formants into a single number. The result cannot be anything but an approximate truth, but it is still true enough to be novel because of its explanatory power in child language development, metathesis, haplology, and sonority scales (see Appendix V). It can even be used to understand how the phonemes of languages are distributed over this volume and what we should expect.

The X-axis (place) takes care of both height and length (in a multiplicative way). The fact that the place of constriction has an effect on the acoustic output is undeniable. Looking at it from the point of view of the two-tube model, the place of constriction changes the size of both tubes. Looking at it from the point of view of the source & filter model, it is obviously the mechanism of changing the parameters of the filter that shapes the acoustic output. Looking at it from the point of view of experiments, the data of Stevens and Blumstein (1978) show that and there are cues for the place of articulation in stop consonants. The role of the amplitude of the fricative noise in the perception of place of fricatives can be seen in Behrens and Blumstein (1988). The role of onset spectra (for example Blumstein & Stevens, 1980) for the stops, and a lot of other experimental results indicate that there does indeed exist acoustic correlates of articulation. Of course, for the vowels, the role of the place of constriction in shaping the peaks of the output is clear and the three formants form a left-handed vector space as can be seen chapters III and IV of Hubey (1994, 1996). And if there still exist other acoustic invariants which have not yet been found, it does not mean that they do not exist. I think I found some which can be shown on this space in conjunction with the vowel (formant) space, but already implies that

we need six dimensions. It will be shown later, that we need yet more dimensions.

The Y-axis definition quite obviously is based on articulation and the acoustic correlates are quite easily found. The stops would probably be best modeled mathematically as the Dirac delta-function. It is well-known that the power spectral density of the delta function is white noise (flat across the spectrum, i.e., constant over all frequencies). Of course, it is just as well-known that no real signal can have power at all frequencies since it would require all energy in the universe, but the approximation works anyway – both in mathematics and physics. In practice, in mathematical modeling and physics 'fat' Gaussians (i.e., large second central moment) are used for white noise. The voiceless stops then would have power at all the frequencies (as a mathematical idealization of course). In practice, we cannot see noise at all frequencies; they would have to drop off with frequency. This is indeed reasonably well-corroborated in the power spectra of the voiceless plosives/stops from Edwards (1992). The slight differences among them are no doubt due to the place of the stop as already mentioned above, noted by Stevens and coworkers. Furthermore, the filter (vocal tract) also acts on the friction noise and shapes the output. The high frequencies decay more readily and all of this can be seen in the power spectra (and is known from the study of wave propagation and communication via electrical signals).

Now, the voiced plosives will have a more complicated spectral density since the output is a result not only of periodicity but also white noise (idealization for friction noise). And this is also easily corroborated (see Edwards, 1992). The spectrum of the voiced plosives is jagged in all cases due to the interaction of voicing (periodic/resonant sounds) with turbulence/chaos in flow. Any plosive will (theoretically) have a spectrum which includes wideband noise since if the 'plosion' is modeled as a highly peaked (i.e., short duration) pulse, then its Fourier transform is constant. In the simplest approximation of this as a Dirac delta function, we can see that the power spectrum is constant. In addition, if the plosive is between two vowels the sample will either contain only the plosion (i.e., the delta function) if the sampling duration is very short, or will contain samples from the surrounding vowels, and hence contain resonances (i.e., high peaks which in clearly identifiable form would be called vowels) which could characterize both or either vowel depending on the sample. Since for voiced plosives, by definition, the voicing is not turned off, then we can expect to see the plosion superimposed (not necessarily linearly) on top of a carrier which seems to be comprised of two vowels. This is what can be seen in Edwards (1992) for all voiced plosives. On the other hand, if the sample is examined in, say, three approximate pieces, the first leading to the plosion containing traces of the preceding vowel, followed by the plosion, and then the last approximate third containing traces of the following vowel in anticipation (i.e., co-articulation) then we would need to examine the locus equations, which is not done here explicitly, although alluded to in several places such as in section 8 (please see Appendix V). However, it is obvious from the speech space that such a VCV would constitute a path/trajectory in the phase space and we can (theoretically) envision this motion in terms of the characteristics of the power spectra as well as in terms of the motion of the articulators (see also Appendix V). Therefore it is approximately true that the VCV path in the phase space will move from highly resonant (low Z value) region towards a momentarily turbulent region (mid Z) and then back towards the high resonance reregion. Therefore, the phase space contains not only acoustic but also articulatory and perceptual significance, as well as being the natural space for the 'locus equation' approach.

The differences in the rates of closure for the voiceless or voiced stops is also known (Allen & Norwood, 1988; Flege, 1988; etc.). Thus the placings of the voiced and voiceless stops/plosives in the phase space are motivated by experimental evidence. The spectral densities do not show the duration, hence cannot show that the spectrum of a voiceless plosive is of very short duration (like a shock wave) and that the voiceless fricatives are steady-state. Essentially, the duration is taken care of indirectly in this space

since the Y-axis is defined as a derivative (more like an impulse because of the Dirac function approximation) but this mathematical model translates directly into the acoustic domain as described above since its acoustic correlate is indeed white noise (or friction noise). The placement of the plosives/stops is not an accident, and the dual nature of the Y-axis is also clear. The only reason an average like the rms is suggested is to take care of the case in which the articulators (say the lips) go through one complete cycle so that the derivative over the cycle would result in zero, if the rms value was not used. It is clear that more dimensions have to be added to be able to handle geminates, ejectives, trills voiceless vowels etc. Only three dimensions were chosen to make use of our human intuitive grasp of three dimensional spaces. The usefulness of this space can easily be expanded by extending it to 7-8 dimensions. Released/unreleased distinctions are already treated since the Y-axis is the absolute value of some kind of a weighted average of the magnitudes. The released/unreleased opening/closing etc. are lumped into a single dimension only along the positive semi-infinite axis. Separating them will have to be accomplished by probably going to complex numbers. Nasality is also a problem like making the [s] & [š] and [z] & [ž] distinction. Nasality would require another dimension, say for the *nasal murmur*. The differences in the placements of these, as can be seen from spectral densities in Edwards (1992) is that they essentially have to do with the shape of the spectrum.

The Z-axis is essentially binary (or at best ternary, considering the transition). It divides the Z-axis into resonant (peaked, or compact) and noisy (flat, or diffuse) spectral shapes. The transition zone has both peakedness and noise. These have been forcefully collapsed into a single dimension in order to avoid having to draw high dimensional spaces on two-dimensional paper. In reality, they would be much better represented as separate dimensions. The mathematical definitions of the dimensions are inadequate for the task; they would have to be slightly modified to take these into account. The place of maximum power has been used as a secondary consideration (as a part of the interpretation). Since the Z-axis already divides the phones into resonant (peaked) and noisy (flat spectrum) and since it just so happens that the vowels have their energy towards the lower frequencies, then for the fricatives and stops, the place of the maximum peak of fricative power was used to distinguish between [s] and [š] and [z] and [ž]. The extra information was used to modify the definitions of the dimensions, which is similar to the idea of Fant in which the information from the third formants was used to modify the second formant so that the phenomena could be represented in two-dimensional space. The results used in the placements of these are from Stevens (1985) and Delattre et al (1964) and can be seen in Lieberman and Blumstein (1988). Changing (moving) the peakedness (center of the friction energy) from 5 KHz to 2.5 KHz changes the perception from [s] to [š]. Similar problems occur with the type and place of fricative energy; [š] and [s] have greater amplitude than [f] and [θ] (Stevens, 1961). It also depends on the relative magnitude of the friction compared to the signal (i.e., the spectral peaks of the vowel). As another example, the 'explanations' only say that if a certain path in this space gets cut short it passes through another region. One example is that [sy], say in *mission*, if cut short goes through [š] and therefore we get /mišin/ instead of /misyin/. There are many of such examples. In fact, there are much more, [mr] > [mbr], [kt] > [ç]. For example, the abrupt rise in amplitude at the consonant release is perceived as a stop and a gradual rise as a continuant (Shinn & Blumstein, 1984). Changing the friction noise can also result in (s) being perceived as a [ç] (Cutting & Rosner, 1974). Obviously, these must be modeled as 'colored noise' not 'white noise'. That means that the definitions of the X,Y and Z as given in this chapter are idealizations; they can be fixed up with minor twiddling or via extending the 3-D space to five or six, in addition to the at least three dimensions for the vowels (Hubey (1994) or Pullum & Ladusaw (1986) for the standard results from Jones to Ladefoged), so that we are really discussing looking into about ten dimensions.

One of the basic notions that is used, that of *dynamic* and *steady state*, is certainly a simplifi-

cation. So is the concept of vowel vs. consonant. If a two-way split is good, then a four way split is even better. Also, the way of distinguishing vowels and consonants is most certainly easily done using the criteria that are used in this paper. It has to do with making discrete articulations; a vowel can certainly be sustained as long as a human has enough air left in his lungs without making any movements of the articulators. Obviously the initial conditions do not count; the start has to take place somewhere. None of the consonants do not have this property, especially the plosives. Of course, the fricatives are steady state but they also do not display the distinct spectral peaks of vowels. Therefore the sounds are divided into four groups, simply extending the binary *consonant-vowel* distinction to semi-vowels (which already forms a part of linguistics) and to quasiconsonants (essentially a steady state consonant). It is all a simplification along the lines of those made by linguists and scientists for centuries and still being made by them.

As regarding things called degrees-of-freedom (for example, /l/ and /z/) the problem is that the simplification is not referring to these sounds for the discussion of normal speech. Obviously the articulators are in constant motion which is very difficult to describe mathematically. Every phonetics theory deals with simplifications. During the production of say, the /l/, we are not discussing the fact that the tongue can hit different spots for a *clear-l* or a *dark-l*. The fact is, the target articulation that produces the /l/ can be held in a steady-state, just like a vowel articulation. This cannot be done with a stop/plosive, for example, since if the articulation is held, there would be no sound. Even if the tongue moves slightly the essential positions can be held in steady-state and in these positions because of the positioning of the jaw, lips, etc., the sound that comes out in the steady state is most closely related to the sound that comes out in the most neutral position. This can be observed in the spectral density of /v/,/z/, etc. in published works (Edwards, 1992). Obviously, the noise level is high and it is hard to see the spectral peaks but the spectral density decays smoothly and exponentially, unlike fricatives like /s/, or /ʃ/ which are flat or increasing, and unlike the plosives whose spectra are jagged (see, for example, Edwards, 1992). All the voiced non-stop consonants have similar spectra in the sense that the whatever the peakedness (i.e., formant like quality) that exists seems to be exponentially decaying as in the [I] or [Λ], (or the origin vector in the figure in Appendix VI) and the noisy component contains resonance because of the voicing and especially because some of the articulators (aside from the glottis) are set in vibration. This is especially noticeable in /v/ and /z/, where the lips and a part of the tongue, respectively, can be set in vibration. The voiced stops display both white noise (only as an idealized property since it is really as colored noise) and periodicity (resonance), as the spectra is jagged somewhat like an irregular sawtooth curve (Edwards, 1992).

As another simplification which is used to produce coherence, we note that vowels like /v/ and /z/ are what might be called *schwa-colored.* The statement comes directly from the spectral density of /v/, /z/, the source & filter model, the spectral density of the schwa-like sounds (Edwards, 1992), the fact that high frequencies do not travel with the same velocity as low frequency ones (except in simplified linear models in undergraduate texts), and the fact that nonlinear distortion is known to be a chief source of mischief in communication lines. It has to do with the ability to know something about the signal and the process that created it by looking at the power spectral density of the signal. The exponential decay of the spectral density of certain sounds is obvious, as can be seen in Edwards (1992). Comparing this to some others where we do not see it tells us a lot about the filter that produced it. The schwa (and its relative /ʎ/ in American English) displays almost perfect exponential decay (Edwards, 1992). It is not too surprising since all the articulators seem to be in their most neutral position for this phone/phoneme; the tongue passes through this position for back and front vowels and consonants, so in that sense it is like an equilibrium position. All we have to do is to take this spectral density and add some noise and we will see that as the noise level starts increasing the spec-

trum will start to resemble the nonstop voiced fricatives and liquids/nasals like [l], [v], [r], [z], etc. It is just that some of them have less friction noise than others and the slight differences among them are the result of the type and place of constriction.

Because of all these difficulties, it was much easier to use an already existent dimensionless number (Reynolds' number) to represent this dimension than to provide the articulatory and mechanical parameters. In all likelihood, a combination of Reynolds' number along with Strouhal number will provide a better fit with data. And of course, by extending the dimensions to about ten we would have a much better description of the system which, while mathematically more accurate, will lack the intuitive obviousness and attractiveness of the simple idealized three dimensional space presented here. This space would be very difficult to visualize and only statistical tests could determine the placements of the phones. This space similarly has been constructed to provide the simplest coherent mathematical space that can be used as an idealized construct to further develop more complicated, more accurate and more sophisticated spaces and to provide a unified perspective for future directions of research and data collection. There could possibly be a brute force approach that could yield better results using dimensionless numbers. It might even produce more dimensionless numbers that could be useful for speech. However, it is likely that the dimensionless numbers of fluid mechanics such as Froude, Euler, Weber, Mach, Prandtl, Eckert, Grashof would somehow show up. The fact that the Strouhal number has to do with oscillating flow could make it useful in conjunction with Reynolds' number for the Z-axis. The X- and Y-axes could certainly be improved upon and only experiments can decide the exact shape of the space. We might try some kind of a combination of the Reynolds' number and Strouhal number for the air quality axis of the space. The alternative is to keep plotting variables against time until doomsday; that will not accomplish anything. There are very fundamental concepts in physics. One of them is that any equation must be dimensionally homogeneous. That is the reason for the power of dimensional analysis; it produces combinations of parameters among which experimental relationships must be sought.

It is the very complexity of the fluid phenomena that makes dimensional analysis so powerful and useful in that field. The fact is that the space that is produced by using dimensionless numbers is able to represent very complex phenomena in a very simple and intuitive way. It is the very nonlinearity that is introduced by using multiplications and divisions of the parameters that unskews the complex web of tangled fluid phenomena and allows the experimental physicists and hydrodynamicists to fit 'nice' curves into their experiments. It is this amazing power of dimensional analysis that allows this simple space to have both *articulatory* and *acoustic* interpretations. This simple linear three dimensional speech phase space can be used to unify many stylized facts of linguistics, as shown.

REFERENCES

Abry, C., Boe, L. & Schwartz J. (1989). Plateaus, catastrophes and the structuring of vowel sytems, *Journal of Phonetics 17,* 47–54.

Allen, G. & J. Norwood, J. (1988). Cues for intervocalic /t/ and /d/ in children and adults, *JASA 84,* 868–875.

Anderson, S. (1985). *Phonology in the Twentieth Century,* Univ. of Chicago Press.

Anderson, S. (1974). *The Organization of Phonology,* Academic Press, New York.

Anderson S. & Ewen, C. (1987). *Principles of Dependency Phonology*, Cambridge, Cambridge University Press.

Assmann, P. & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundemantal frequencies, *JASA 88,* 680–697.

Atal, B., & Schroeder, M. (1978). Linear prediction analysis of speech based on a pole-zero representation, *JASA 64,* 1310–1318.

Bailly, G., Laboissiere, R., & Schwartz, J. (1991). Formant trajectories as audible gestures: An alternative for speech synthesis, *Journal of Phonetics 19,* 9–23.

Banks, S. (1990). *Signal Processing, Image Processing and Pattern Recognition,* Prentice-Hall.

Behrens, S. Blumstein, S. (1988). On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants, *JASA 84,* 861–866.

Brainerd, B. (1971). *Introduction to the Mathematics of Language Study,* Elsevier, New York.

Broad, D. & Wakita, H. (1977). Piecewise planar representation of vowel formant frequencies, *JASA 62,* 1467–1473.

Browman, C. & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures, *Journal of Phonetics 18,* 299–320.

Carre, R. & Mrayati, M. (1991). Vowel-vowel trajectories and region modeling, *Journal of Phonetics 19,* 433–443.

Catford, J.C. (1988). *Phonetics,* Clarendon Press, London.

Chiba, T. & Kajiyama, J. (1941). *The Vowel: Its nature and its structure,* Tokyo,Tokyo-Kaiseikan Publishing Company.

Clark, J. & Yallop, C. (1990). *Phonetics and Phonology,* Blackwell, Oxford.

Damasio, A. & Damassio, H. (1992). *Brain and Language,* Scientific American, p. 88.

Dayhoff, J., (1990). *Neural Network Architectures,* Van Nostrand Reinhold.

Edwards, H. (1992). *Applied Phonetics: The Sounds of American English,* Singular Pub.

Fant, G. (1990). *Acoustic Theory of Speech Production,* Blackwell, Oxford.

Fischer, R. & Ohde, R. (1990). Spectral and duration properties of front vowels as cues to final stop-consonant voicing, *JASA 88,* 1250–1259.

Flanagan, J. (1972). *Speech analysis, synthesis and perception,* New York, Springer-Verlag.

Flege, E. (1988). The Development of skill in producing word-final English stops: Kinematic parameters, *JASA 84,* 1639–1652.

Foley, J. (1977). *Foundations of Theoretical Phonology,* Cambridge Univ. Press, Cambridge.

Fromkin, V. (ed), (1985). *Phonetic Linguistics: Essays in Honor of Peter Ladefoged,* Academic Press, New York.

Ganong, W. & Zatorre, R.(1980). Measuring phoneme boundaries four ways, *JASA 68,* 431–439.

Goldsmith, J. (1990). *Autosegmental and Metrical Phonology,* Blackwell, Oxford.

Gopal, H. (1990). Effects of speaking rate on the behavior of tense and lax vowel durations, *Journal of Phonetics 18,* 497–518.

Greenberg, J. (ed.) (1978). *Universals of Human Language: Phonology,* Stanford University Press, Standford.

Hankamer, J., Lahiri, A. & Koreman, J. (1989). Perception of consonant length: voiceless stops in Turkish and Bengali, *Journal of Phonetics 17,* 283–298.

Hawkins, J. (ed.) (1988). *Explaining Language Universals,* Basil Blackwell, New York.

Henton, C. (1990). One vowel's life (and death ?) across languages: the moribundity and prestige of /L/, *Journal of Phonetics 18,* 203–227.

Hillenbrand, J., Canter, G. & Smith, B. (1990). Perception of intraphonemic differences by phoneticians, musicians and inexperienced listeners, *JASA 88,* 655–662.

Hubey, H.M. (1993). Psycho-socio-economic evolution of systems, in *Mathematical Modelling and Scientific Computing,* (X. Avula, ed.), Principia Scientia, St. Louis.

Hubey, H.M. (1996a). Catastrophe theory and speech, submitted to *Journal of Nonlinear Dynamics, Psychology, and Life Sciences.*

Hubey, H.M., (1996b). Speech realization, fuzzy sets, differential equations and categorical perception, submitted to *Journal of Nonlinear Dynamics, Psychology, and Life Sciences.*

Hubey, H.M. (1999). *The Diagonal Infinity,* World-Scientific, Singapore.

Hubey, H.M. (2000). *Mathematical and Computational Linguistics,* to be published by Lincolm-Europe. First edition published in 1994, Mir Domu Tvoemu, Moscow, Russia.

Hyman, L. (1975). *Phonology: Theory and Analysis,* Holt, Rinehart and Winston, New York.

Jackson, E. (1991). *Perspectives on nonlinear dynamics 1,* Cambridge University Press, Cambridge.

Jakobson, R. (1990) *On Language,* Harvard University Press, Cambridge, MA.

Kelso, J., Saltzman, E. & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory, *Journal of Phonetics, 14,* 29–59.

Kelso, J., Saltzman, E. & Tuller, B. (1986). Intentional contents, communicative context, and task dynamics: A reply to commentators, *Journal of Phonetics, 14,* 171–196.

Kewley-Port, D. & Atal, D. (1980). Perceptual differences between vowels located in limited phonetic space, *JASA 85,* 1726–1740.

Klatt, D. (1980). Software for a Cascade/Parallel Formant Synthesizer, *JASA 67,* 971–995.

Ladefoged, Peter (1975). *A Course in Phonetics,* Harcourt, Brace & Jovanovic.

Ladefoged, Peter (1962). *Elements of Acoustic Phonetics,* Univ. of Chicago Press, Chicago.

Ladefoged, Peter (1971). *Preliminaries to Linguistic Phonetics,* Univ. of Chicago Press, Chicago.

Ladefoged, P. & Maddieson, I. (1990). Vowels of the world's languages, *Journal of Phonetics 18,* 93–122.

Ladefoged, P. (1990). Some reflections on the IPA, *Journal of Phonetics 18,* 335–346.

Lass, Roger (1984). *Phonology,* Cambridge University Press, Cambridge.

Lieberman, P. (1984). *The Biology and Evolution of Language,* Harvard University Press, Cambridge.

Lieberman, P. & Blumstein, S. (1988). *Speech Physiology, Speech Perception and Acoustic Phonetics,* Cambridge University Press, Cambridge.

Lindblom, B. (1990). On the notion of "possible speech sound", *Journal of Phonetics 18,* 135–152.

Lindblom, B. & Engstrand, O. (1989). In what sense is speech quantal? *Journal of Phonetics 17,* 107–121.

Lindblom, B. & MacNeilage, P. (1986). Action Theory: Problems and Alternative Approaches, *Journal of Phonetics, 14,* 117–132.

Lubker, J. (1986). Articulatory Timing and the Concept of Phase, *Journal of Phonetics, 14,* 133–137.

Lyons, J. (1968). *Theoretical Linguistics,* Cambridge University Press, Cambridge.

Makkai, V. (ed.) (1972). *Phonological Theory,* Holt, Rinehart and Winston, New York,

Manuel, S. (1990). The role of contrast in limiting vow-

el-to-vowel coarticulation in different languages, *JASA 88,* 1286–1298.
Miller, J. (1989). Auditory-perceptual interpretation of the vowel, *JASA 85,* 2114–2133.
Moore, B., Peters, R., & Glasberg, B. (1990). Auditory filter shapes at low center frequencies, *JASA 88,* 132–140.
Mowrey, R. & MacKay, I. (1990). Phonological primitives: Electromyographic speech error evidence, *JASA 88,* 1299–1312.
Nearey, T. (1978). *Phonetic features for vowels,* Indiana Univeristy Linguistics Club, Bloomington, Indiana,
Nearey, T. (1990). The segment as a unit of speech perception, *Journal of Phonetics 18,* 347–373.
Ohala, J. & Jaeger, J. (1986). *Experimental Phonology,* Academic Press, New York.
Ohala, J. (1990). There is no interface between phonology and phonetics: a personal view, *Journal of Phonetics 18,* 153–171.
Ohala, J. (1986). Against the direct realist view of speech perception, *Journal of Phonetics, 14,* 75–82.
Peterson, G. & Barney, H. (1952). Control methods used in the study of vowels, *Journal of the Acoustical Society of America, 24,* 175–184.
Picone, J. (1990). Continuous speech recognition using hidden Markov models, *IEEE ASSP Magazine.*
Repp. B. (1986). Perception of the (m)- (n) distinction in CV syllables, *JASA 79,* 1987–1999.
Saporta, S. & Bastian, J. (eds.) (1961). *Psycholinguistics: A Book of Readings,* Holt, Rinehart and Winston, New York.
Schroeder, M. (1973). An integrable model for the basilar membrane, *JASA 53,* 429–434.
Secker-Walker, H. & Searle, C. (1990). Time-domain analysis of auditory-nerve-fiber firing rates, *JASA 88,* 1427–1436.
Shailer, M., Moore, B. Glasberg, G. & Watson, N. (1990). Auditory filter shapes at 8 KHz and 10 KHz, *JASA 88,* 141–148.
Shinn, P. & Blumstein, S (1983). "Phonetic disintegration in aphasia: Acoustic analysis of spectral characteristics for place of articulation," *Brain and Language 20,* 90-114.
Shirai, K. & Kobayashi, T. (1991). Estimation of articulatory motion using neural networks, *Journal of Phonetics 19,* 379–385.
Silverman, H. & Morgan, D. (1990). The application of dynamic programming to connected speech recognition, *IEEE ASSP Magazine.*
Sinha, N. & Kuszta, B. (1983). *Modeling and Identification of Dynamic Systems,* Van Nostrand Reinhold, New York.
Siski, R. (1967). Stochastic differential equations in *Modern Nonlinear Equations* (T. Saaty, ed.), McGraw-Hill, New York.
Sondhi, M. & Gopinath, B. (1971). Determination of vocal-tract shape from impulse response at the lips, *JASA 49,* 1867–1873
Stevens, K. & Blumstein, S. (1978). Invariant Cues for Place of Articulation in Stop Consonants, *JASA 64,* 1358–1368.
Stevens, K. (1989). On the quantal nature of speech, *Journal of Phonetics 17,* 3–45.
Strange, W. (1989). Evolving theories of vowel perception, *JASA 85,* 2081–2087.
Sundberg, J. & Lindblom, B. (1990). Acoustic estimation of the front cavity in apical stops, *JASA 88,* 1313–1317.
Sussman, H. (1990). Acoustic correlates of the front/back vowel distinction: A comparison of transition onset versus "steady state", *JASA 88,* 87–96.
Syrdal, A. & Gopal, H. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels, *JASA 79,* 1086–1100.
ten Bosch, L. & Pols, L. (1989). On the necessity of quantal assumptions. Questions to the quantal theory, *Journal of Phonetics 17,* 63–70.
Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale, *JASA 88,* 97–100.
Treiman, R., Gross, J. & Cwikiel-Glavin, A. (1992). The syllabification of /s/ clusters in English, *Journal of Phonetics 20,* 383–402.
Visch, E. (1990). *A Metrical Theory of Rhythmic Stress Phenomena,* Doris Pub., Providence.
Van Son, R. & Pols, L. (1990). Formant Frequencies of Dutch Vowels in a Text, Read at Normal and Fast Rate, *JASA 88,* 1683–867.
Waibel, A. & Lee, K. (Ed) (1990). *Readings in Speech Recognition,* Morgan Kaufman Publishers, San Mateo, CA.
White, F. (1979). *Fluid Mechanics,* McGraw-Hill.
Witten, I. (1982). *Principles of Computer Speech,* Academic Press, New York.
Winter, I. & Palmer, A. (1990). Temporal responses of primary-like anteroventral cochlear nuclear units to the steady-state vowel /i/, *JASA 88,* 1437–1441.
Zwicker, E. & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frequency, *JASA 68,* 1523–1524.

APPENDIX I

Formants and Noise

One of the first things we notice is that the formants for the neutral vowel [s] (schwa in short form in English, or a related one [Λ], and as [ɪ] in Turkish (often depicted as /ï/)) have more regularity and symmetry about them than the other vowels. They are more evenly spaced and the amplitudes drop off exponentially. Therefore, the filter (the articulator apparatus) that produces this shape is more natural in that sense. We normally expect higher attenation at higher frequencies. In that sense the formant patterns for the other vowels may be thought of as implementing what is called pulse-position coding (PPM) as done in digital encoding of analog signals for transmission. Although the most commonly used method now is PCM (pulse code modulation), PPM was used experimentally. In the PPM coding it is the position of the pulse relative to where it should have been compared to a regular clocking signal that was indicative of the value which PPM encoded. In that sense, the perception of vowels may be in the relative positions of the formants which means that the vowels are recognized from their patterns instead of from their absolute positions in the spectrum. See Hubey (1994) and Nearey (1978).

It is only in the power spectra of the vowels that we can notice such distinct peaks. The consonants have spectra distinctive of their type. For example, the fricatives have much noise added to the power spectrum (which is indicative of the background vowel which is being modulated). The quasiconsonants in this sense are phonemes which have formant structures which resemble the neutral vowel but the friction noise is added. In analytical studies of stochastic processes the noise is often added (sometimes multiplied) to the deterministic process which in physical problems is usually described by a differential equation. For analytical tractability instead of white noise (equal power at all frequencies) one often uses Gaussian noise with a large variance or dispersion. In reality, white noise cannot exist since such noise would have infinite power. In modeling of friction noise in linguistics one can add almost constant power noise at the lower frequencies as can be seen from the power spectra of fricatives such as /s/, /š/, /f/. The noise is not really constant but, in general, rather slowly tends to fall off towards the higher frequencies. In addition, the position of where most of the energy is concentrated in friction noise signals what phoneme will be perceived as has been shown by Stevens and coworkers. If we look at typical power spectra of the quasiconsonants we see that we can obtain such a power spectrum by adding friction noise to the neutral vowel spectrum, as can be seen in the figures in which more and more noise is added to the /ɪ/. The truth of these general ideas is abundantly clear from the power spectra of various speech sounds, for example, for English (Edwards, 1992).

A new process can be created from an old one by

$$y(\omega) = H(S(\omega) - S_0) \tag{I.1}$$

where H (x) is the unit step function, x_0 is some threshhold and which may be some function of the mean of the process, say, $S_0 = \frac{\alpha}{\Omega}\int_0^{\Omega} S(\omega)\, d\omega$.

Then y (ω) is a reasonable measure of the peakedness of the smoothed power spectrum of the phone/phoneme. The formal derivative is

$$\dot{y}(\omega) = \dot{S}(\omega)\, \delta[S(\omega) - S_o] \tag{I.2}$$

where δ (x) is the Dirac delta function. We can obtain the number of crossings of the threshhold by the power spectrum which is a measure of the peakedness. The number of times the spectrum crosses the threshhold in the interval (0,Ω) is then

$$N[S_0;0,\Omega] = \int_0^{\Omega} |\dot{S}(\omega)|\, \delta[S(\omega) - S_o] \tag{I.3}$$

In addition, we would probably like to multiply this by the difference in the maximum and minimum values since it is in the peaked spectra of vowels that we see great variations and it is in the spectra of voiceless fricatives that we note a flattish power spectrum. This product will also make it easier to distinguish the jagged spectra

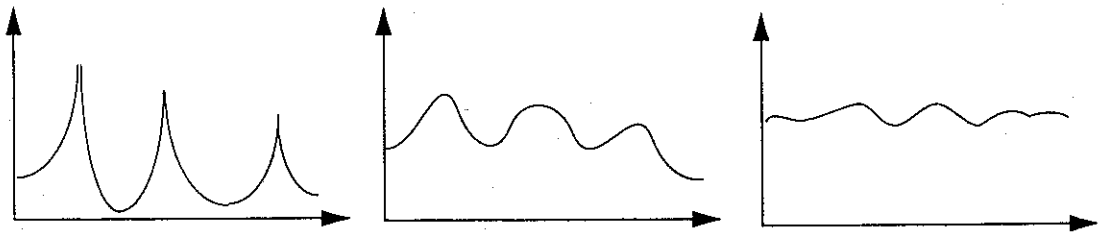

Fig. I.1. Peaked and flat spectra and noise (frication). It can be seen that the peaked spectrum of vowels is different than the flattish spectra of semivowels, and glides. Furthermore, the amount of noise can bury much of the spectral information as can be seen on the extreme right. These spectra should be compared to typical spectra of vowels and nonvowels in Edwards (1992).

that one observes for the voiced plosives and the vowel-like spectra one observes for the liquids and nasals. Therefore we can define the 'peakedness' as

$$P(\omega) = [max\ (S\ (\omega)) - min\ (S\ (\omega))]N[S;(0,\Omega)] \quad (I.4)$$

If the spectrum has been smoothed, i.e., we have a spectrogram, then the maximum and minimum of the spectrum will occur where the derivative is zero. We can see that P (ω) can distinguish between these three types of spectra

APPENDIX II

High and Low Resolution Descriptions

Phone is a speech sound. A phoneme is a minimal set of phones of a given language which serves to distinguish between at least two words. Therefore, from the definitions we can see the need for yet another concept: a phonete. We can consider these from two related points of view; one from fuzzy sets and fuzzy logic, and the other from stochatic processes.

If we could find some appropriate phase space in which all speech sounds of humans can be represented then each point in this space would represent a phone. Conversely every word uttered by anyone speaking any language consists of phones. In this sense phones are merely sample functions being produced by some deterministic process (i.e., the speech producing articulation mechanism). Therefore, phones are instantiations (particular manifestations, particular instances, or sample functions) of phonemes. Or in another way, phonemes may be considered to be the best representatives of fuzzy sets of phones. Or yet another way, the phonemes are volumes in the phase space (fuzzy volumes) which contain particular instances (phones) of their manifestation. In a simplified way, phonetes are high-resolution generic representations of speech sounds, whereas phonemes are low-resolution representations of particular sets of sounds specific to given languages.

Obviously we still have a problem in that if the phase space of speech sounds is universal in the sense that all human speech sounds must be representable in this space, although we can represent every speech sound of every language, we recognize that the specific ways in which this hypervolume is divided up into distinctive speech sound sets of languages varies from one language to another. In this case we still cannot discuss speech in all generality except by creating yet more subsets or hypervolumes in this space which are still generic, and universal but lower resolution than phonetes. Similar problems in translation of natural languages, compilation of programs, are handled by converting all to a standardized code or language. In the case of the real world, both using actual languages (say English presently) or artificially constructed language (such as Esperanto) have been tried. We can use a similar idea in thinking of representative sounds common to many languages but exemplified in terms of sounds or

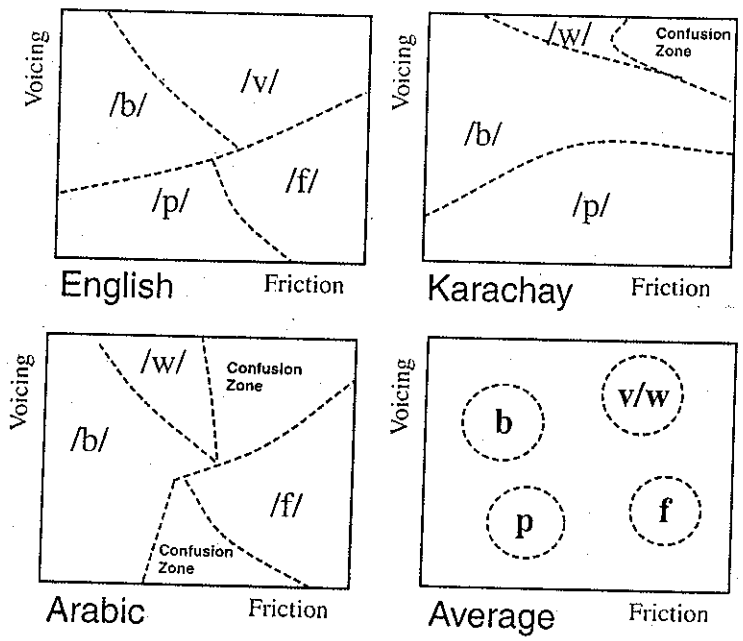

Fig. II.1. A highly suggestive view of the possibility of dividing up the available space for speech sounds.

phonemes of actual languages to provide a frame of reference. Therefore the symbols used are taken from common usage. Some are IPA symbols and others are commonly used symbols which may differ slightly from IPA usage. It is not necessary to use IPA symbols because the discussion is not in terms of high resolution objects but are somewhere in between phones and phonemes. The circles in Figures 5 and 6 represent the best representative of a particular class of speech sounds which can be recognized to be such in all the languages in which it exists. For example, in two dimensions one may see that a particular area (two-dimensional hypervolume) may be divided up in several ways. For example, along the two dimensions of voicing and friction (ignoring the other articulatory organs) we might divide up the available space as in English, Arabic or Karachaic (a language of the North Caucasus). Although English has all four phonemes, /b/,/p/, /f/, and /v/, Arabic lacks a /v/ and a /p/. For example, the Turkish /paşa/ has made its way into Arabic as /başa/. Similarly, Karachaic has no /v/ or /f/ (although Anatolian Turkish has picked up both from Arabic and Farsi) so that words such as /fatima/ are /patimat/ or /baydımat/. There is a need to represent these sounds not as phonemes (which are relative to given languages) and not as phones (because there are an infinite number or phones), but as sets of sounds similar to phonemes but as generic examples; phonetes. We might do this by using the fuzzy intersection of all such sounds that are recognizably similar (by trained individuals), or we can represent them as small circles which may be thought of as the best representatives of such sounds. This idea is not too far from fuzzy logic since one of the ways of defining objects difficult to define in fuzzy reasoning is by producing best examples. For example, the best examples of 'fruit' might be 'apple', followed by 'orange', etc. We can assign values in [0,1] to the 'fruitness' or 'fruitiness' properties of specific examples of the set of fruits, such as assigning 0.9 to apple or 0.8 to orange.

In this case the symbols used do not necessarily have to be the phonetic symbols. In the case of selecting a point in a vector space as the best

representative (or switching to the stochastic view the center or average of a distribution of phones) the use of vector notation is quite natural instead of the symbols for phonemes such as the two slashes /./ or the phonete symbol as [.]. Quite naturally, as a practical matter, the phonetic symbols are relics of the mechanical printing press era. There is presently no difficulty of finding or producing many different symbols on sophisticated word processors. As a matter of future importance, since Unicode (a 16 bit code which is sufficient for 65,536 symbols) is spreading along with Java which is a de facto standard for the World Wide Web (WWW), there is really no reason to keep some of the IPA symbols such as colon, macron etc. Some of the symbols used in this paper which depart from the IPA are {z} for j or zh as in 'measure'; {s} for sh or ʃ as in shirt; ç for tʃ as in chip, c for dz as in gyp; {ð} as in then; θ for th as in thin; **I** as a vector representing a sound very close to English schwa, or ʌ as in hut, or I as in it; ɛ as in head; ü as in German and Turkish; ö as in German and Turkish and close to the sound in 'herd' in English (but without the rotacization i.e., r-coloring).

These ideas were expressed as early as 1950 as can be found in Joos and Hockett (see for example Saporta & Bastian, 1961). Hockett writes, in his review of Shannon and Weaver's book on Information Theory (Saporta & Bastian, p. 51):

> The acoustician examines speech signals and reports that they are continuous. The linguist examines them and reports that they are discrete. Each uses operationally valid methods, so that both reports must be accepted as valid within the limits defined by the operations used, and the apparent contradictions between the reports constitutes a real, not an imaginary problem... The linguist... also, is unable to to examine the speech-signal directly. The ear and the associated tracts of the central nervous system constitute a transducer of largely unknown characteristics...
> A continuum can be transformed into a discrete sequence by any of various QUANTIZING operations; ...though the quantizing operations used in electronic communications are all quite arbitrary. Similarly, a discrete sequence can be transformed into a continuum by what might be called a CONTINUIZING operation. Now if the continuum-report of the acoustician and the discrete-report of the linguist are both correct, then there must be, for any given body of raw material, a quantizing operation which will convert the acoustician's description of the raw material into that of the linguist, and a continuizing operation which will do the reverse; the desired quantizing and continuizing operations must be inverses of each other.

In the same paper a very beautiful description of a stochastic process, attributed to Joos by Hockett, is given:

> Let us agree to neglect the least important features of speech sound, so that at any moment we can describe it sufficiently well with n measurements, a point in n dimensional continuous space, n being not only finite but also fairly small, say six... Now the quality of sound becomes a point which moves continuously in this 6-space, sometimes faster and sometimes slower, so that it spends more or less time in different regions, or visits a certain region more or less often. In the long run, then, we get a probability density for the presence of the moving point anywhere in the 6-space. This probability density varies continuously all over the space. Now wherever (one) ...find a local maximum of probability density, there the linguist finds an allophone; and 'there will be not only a finite but a fairly small number of such points, say less than a hundred.

These descriptions should be compared to the three-dimensional phase space for speech. It is not yet clear how many of these 'local maximum probability densities' exist in languages. Introspection gives one set of answers and speech recognition researchers give another set. For example, we find in Clark and Yallop (1990) that there are 43 English phonemes (21 vocalic and 24 consonantal) which is more than that giv-

en in Chomsky (Speech Patterns in English); Kai-Fu Lee uses 48 phonemes in their Hidden Markov Models of speech recognition (Waibel & Lee, 1990:352); whereas Roucos and Dunham claim that their model uses 270 phonemes (Waibel & Lee, 1990:369); Churchland categorically states that English has 79 phonemes (Forrest, 1991:285); and Ladefoged gives evidence from Moskowitz, Ohala and Jaeger that 'people can use orthographic knowledge as the basis for forming phonological classes' (Dressler et al., 1988:166). In this paper all of the arguments are moot because neither phones nor phonemes have been used, although if phonemes are used they are with respect to a given language, and if phones are used they are in English (and refer to specific references).

APPENDIX III

Wave Equation, and the Source & Filter Model

The equation for wave propagation (including sound/acoustic ones) is given (without coefficients) by

$$\{\partial^2/\partial t^2 + \partial/\partial t - \nabla^2\}\Psi(\mathbf{r},t) = \delta(\mathbf{r}-\rho)\delta(t-\tau) \quad \text{(III.1)}$$

The Fourier Transform with respect to the space variable results in the damped harmonic oscillator equation, the one dimensional version of which is given by

$$\{d/dt^2 + 2\xi\omega_n d/dt + \omega^2{}_n\}\Psi(t) = s(t) \quad \text{(III.2)}$$

The homegenous solution for the damping $\xi < 1$ is given by

$$\Psi(t) = C_1 e^{(-\zeta t - i\Omega t)} + C_2 e^{(-\zeta t + i\Omega t)} \quad \text{(III.3)}$$

where $\Omega = \omega_n (1-\xi^2)^{1/2}$ is the damped frequency of oscillation, and $\xi\omega_n$ is the damping. If we take a temporal Fourier Transform of equation (2) with the Dirac delta function $\delta(t-\tau)$ in place of the source/forcing term we obtain

$$-\omega^2 g(\omega) + 2i\xi\omega_n\omega g(\omega) + \omega_n{}^2 g(\omega) = 1 \quad \text{(III.4)}$$

where $g(\omega)$ is the Fourier Transform of the Green's function of the differential equation. From this we can see that

$$g(\omega) = 1/(-\omega^2 + 2i\xi\omega_n\omega + \omega_n^2) \quad \text{(III.5)}$$

The frequency response of the differential equation, or the spectral density of the response is $|g(\omega)|^2$ which is plotted in Figure (III.1) for different values of the damping ξ for arbitrary values of the ordinate. The natural frequency $\omega_n=3$ as can be seen in the plot. As can be seen, the response is greater for smaller damping. The plots are for $0 < \xi < 1$. Because of the effect of this shape in changing the forcing (input) to the output (response) the transform of the Green's function is called the 'filter'. As can be seen this is the 'filter shape' since the response of the system to various frequencies is given in these plots. The resonance occurs at the source frequency that equals the natural frequency of the system. The source-filter model of speech is essentially about this except that it is a more complex version. Finally, as can be seen if we use three of these filters of different natural frequen-

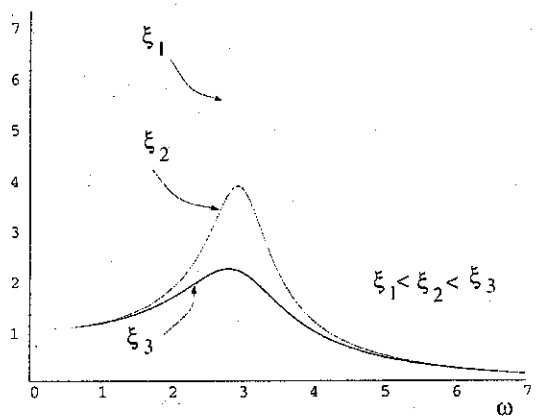

Fig. III.1. Spectral response of the damped harmonic oscillator: This response pattern is used for speech synthesis both in analog and in digital form.

cies, then we can model the first three formants of the vowels of human speech sounds. The two-tube filter (model) of sound production refers essentially a solution of the wave equation with the required boundary conditions and which is a more complex case of the above derivation. Since the RLC circuits of electromagnetics obey the same equation as Eq. (III.2), then analog methods can be used as resonators to produce artificial speech sounds. Finally, digital speech synthesis, such as MITalk, employs the digital computer version of the same filter above.

The Green's function g (t-τ) of the differential equation given in (III.2) is the inverse Fourier Transform of Eq. (III.5). Once we have g (t-τ) we can then express the solution of (III.2) to any kind of source or forcing as

$$\Psi (t-\tau) = \int g (t-\tau)\, s (\tau)\, d\tau \qquad \text{(III.6)}$$

so that the term 'source' can be justified in that it is the source (forcing) which drives the sytem and produces the output Ψ (t-τ).

APPENDIX IV

Vectors

There are many different representations of a vector. It is simply an ordered *n-tuple*, and it is called an *array* in computer science. A *feature bundle* is an array or a vector. The number of features is the dimensionality of the vector. As long as the operations on the vector are clearly defined, they can be applied to many different problems in many different representations. Perhaps the simplest way to think of a vector is to imagine it in pictorially as an object that has a *direction* and a *magnitude*. We can then easily represent it as an arrow. The length of the arrow will be called its *magnitude* and its direction is obvious. Figure IV.1.A below shows two vectors **A** and **B**. Both the overbar and bold notations are used to denote vectorial quantities (overbar in the figures and bold in the text). Geometrically, one way to add vectors is to put them head to tail and then draw another vector from the left-over tail to the remaining head of one of the vectors. This is shown in Figure IV.1.A. The sum of the two vectors **A** and **B** is then another vector **C**. Another way to add them geometrically is to put the vectors tail to tail and draw lines parallel to the vectors; the intersection of these lines is the head of the sum of vector **A** and **B** (that is the vector **C**) as can be seen in Figure IV.1.B. Since these are two-dimensional vectors; (that is, it takes two parameters or variables to represent two-dimensional space) we can represent them in algebraic terms by superposing them on say the XY plane, as shown in Figure IV.1.C The endpoints of the vectors are really what we would call the coordinates of a point.

However, points in 2-D have no direction. In order to show this additional property of vectors (i.e., direction) we have to represent them slightly differently. Since, as can be seen in Figures A and B, we can move the vectors around in space without affecting their addition property or their magnitude or direction, we put the vectors with their tails at the origin (Figure IV.1.C). We can then decompose them into their components along the X- and Y-axis. Indeed, since they are two dimensional (i.e., array was another representation), these components constitute the vector. These components are denoted subscripts as can be seen in Figure IV.1.C. The addition of the two vectors **A** and **B** then corresponds to adding up their separate (X and Y) components to yield the X and Y components of the vector **C.** We have to have some notational device to keep the components separate. There are many ways to do this; one way is to represent it as an ordered pair as in $\mathbf{A}= (A_x,A_y)$, $\mathbf{B}= (B_x,B_y)$ and $\mathbf{C}= (C_x,C_y)$. Thus, vector addition rules yield

$$\mathbf{C}= (A_x + B_x, A_y + B_y) \text{ since } C_x = A_x + B_x \text{ and } C_y = A_y + B_y \qquad \text{(IV.1)}$$

The components of vectors are *scalars*, since they possess only the property, magnitude. Another notational device that is often used is that of a concept of unit vectors. These are vectors that point in the direction of X and Y but their magnitudes are unity. Thus, if we use $\mathbf{u}_x$ and $\mathbf{u}_y$ to denote the unit vectors in the X and Y directions, then the vectors **A** and **B** can be written as

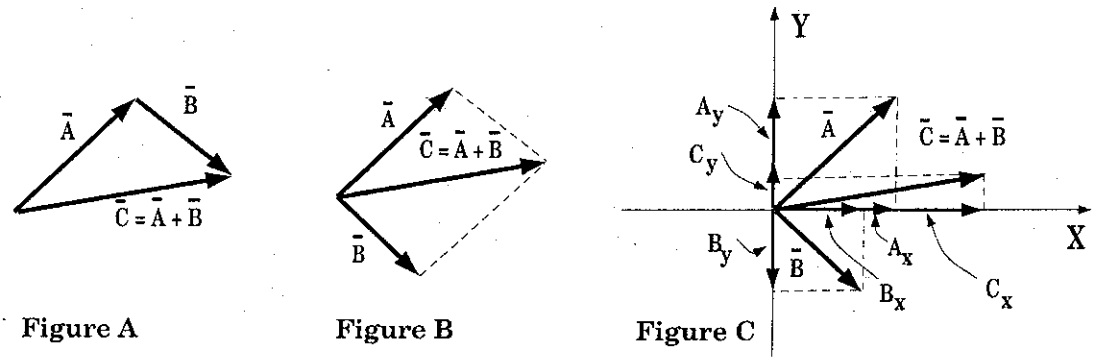

Fig. IV.1. Vectors in two dimensions.

$$\mathbf{A} = A_x\mathbf{u}_x + A_y\mathbf{u}_y ; \mathbf{B} = B_x\mathbf{u}_x + B_y\mathbf{u}_y \text{ and } \mathbf{C} = C_x\mathbf{u}_x + C_y\mathbf{u}_y \quad \text{(IV.2)}$$

Thus we have

$$\mathbf{C} = C_x\mathbf{u}_x + C_y\mathbf{u}_y = \mathbf{A} + \mathbf{B} = (A_x + B_x)\mathbf{u}_x + (A_y + B_y)\mathbf{u}_y \quad \text{(IV.3)}$$

As an application of vectors consider the vectors below in the YZ subspace of the phase space developed. The space is shown below for convenience.

Phonemes or even groups of phonemes are indeed vectors in this space; *fuzzy vectors* but still vectors. We might think of the *centers of gravity* (defined in some weighted sense) of the small volumes in this space to be the unfuzzy vectors that we are discussing in this section. Two vectors **P** and **R** are shown in the figure next to the sonority scale. The vector **P** in the figure below points away from the origin and it can be seen that its components both point in the positive Y and Z direction. Vector **R** points towards the origin and both of its components are negative (that is pointing towards the negative Y and negative Z directions).

APPENDIX V

Path Integrals and Minimization

Many linguistic phenomena can be clearly shown to be the result of some physical optimization effect, that is it can be easily seen to be

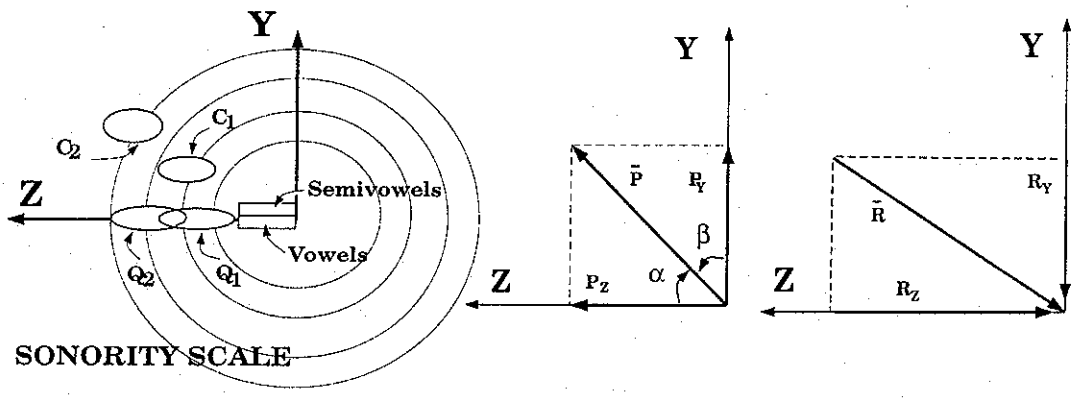

Fig. IV.2. Vectors in sonority space.

minimizing the path length in the *phoneme phase space*. From the figure above we can easily explain the phenomena as *path integral minimization*. For example, the phonetic or acoustic realization of the words; *toes, haws, hods, cleans* is with a /z/, but instead we have *huts, tucks, buts, bits*. It is easy to see why from the diagrams. Another example, *magyar* in Turkish becomes *macar*. The figure shows that /c/ (voiced palatal fricative) is between /g/ and /y/ and the trip from a vowel to /gy/ and then back to a vowel is longer than the trip from vowel to just plain /c/ and back to a vowel. The path from a vowel to /sy/ back to vowel (i.e., *mission*) is long but the /š/ is only part of the way to /s/. The transitions /tb/→/pb/ *(ratbag → rapbag)*, /tm/ →/pm/ *(oatmeal)*, /vt/→/ft/ *(have to)* also can be explained easily in terms of motion in this space. Since the space here symbolizes the motion of the articulatory organs, the distance in this phase space seems to mimic the actual (real) motion of real objects (i.e., articulators) moving in real space possessing momentum and mass. Thus, the /vt/→/ft/ is actually an *overshoot* which can be explained very easily in terms of physical processes such as momentum, inertia, energy and the force required to execute the motions. The other cases were *undershoot*, since it amounted to cutting the path short. One may make an analogy to making turns with a car; at high speeds, tight corners cannot be taken and will overshoot, and at slow speeds, one can make very sharp turns. Others such as /kt/ →/k/ *(facts→ faks)*, /fth/ →/f/ *(fifths →fifs)*, /st/ →/s/ *(chest→chess)* involves cutting the zigzag path short by interpolating the zigzag curves and is the same kind of momentum problem in articulation. More examples of tortuous zigzags that have been smoothed; /tr/→/çr/ *(tree→/çriy/)*, /dr/→/cr/ *(drive→/crayv/)*, *half* but *halves*, *calm* (no /l/), *psalm* (no /p/).

More patterns involving inertia and momentum can be found in *masses, cars, riches, ridges, losses* (all manifesting the ending as /z/). The changes /mb/ → /mbr/, /ml/→/mbl/, and /nb/ → /mb/ can also be seen in terms of the paths in this space. The consonant harmony such as one syllable words having only voiceless plosives such as *pat, pot, cot*, etc. is also explicable in terms of inertia, acceleration and force. The *tenseness* is also easily explained in terms of motion in this space and the duration of the various segments of the path. We can make some general comments about motion in the *phase space*. In so far as it seems to mimic the motion of real articulators in real time-space, we should not expect zigzag paths. If we were to imagine words being constructed as paths in this dimension, we should imagine them as smooth curves since momentum and acceleration effects of

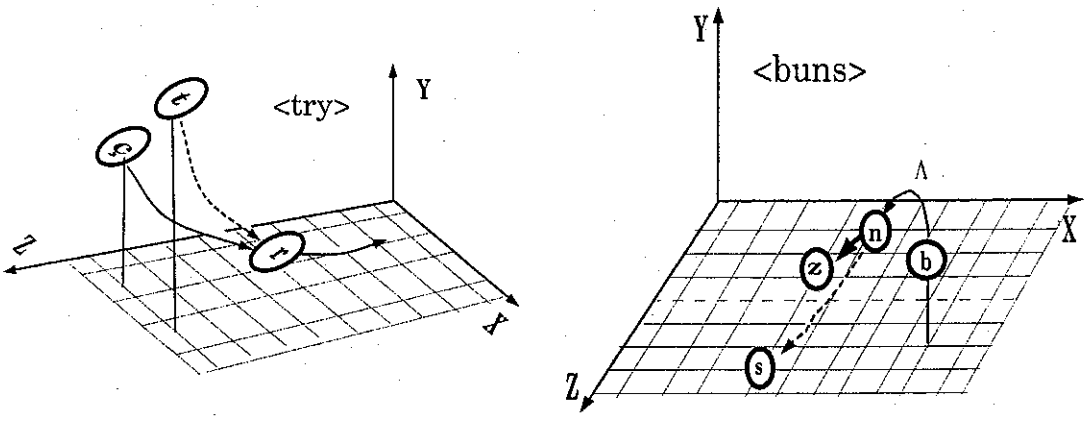

Fig. V.1. Activity in phase space: Because of its dimensions, this phase space is ideally efficient for displaying real world phenomena which has inertia effects such as metathesis, assimilation, haplology.

physics will inhibit sharp motions because of its cost in energy. Tenseness-laxness can also be explained on this basis. Now, if we were to pass smooth curves (such as fitting *cubic splines*) through these points, we will notice that they will tend to be distorted helical shapes. If the turns are very sharp, or if the distances too far, they will tend to get smoothed out. On the other hand, tightly wound curves (such as repetitions) will also tend to stretch out. *Assimilation, metathesis, haplology, dissimilation,* and some of the other linguistic effects can be shown in the phase space to be mostly inertia, acceleration and momentum effects. This statement should not be interpreted to prejudice statements regarding the linguistic disambiguation efforts to place separate semantemes in separate phonological spaces. Thus, if several words (or phonological manifestations of semantemes; that is, *words* or *lexemes*) collide in the higher-dimensional *phase spaces* then there may be efforts to disambiguate even if it means long paths. Of course, these are due to the phonological constraints of languages. Thus, we can think of natural changes occurring in languages due to physiological reasons (ultimately explicable in physics) if not inhibited by the phonological (i.e., phonemic) constraints of languages. Of course, there will be interactions of both physiological and phonological factors. From the previous discussions on the phase space it would be natural to ask if the space is primarily articulatory, acoustic or both. The phase space is both and the dimensions (i.e., the dimensionless groups) can be described in both articulatory and acoustic terms. It shows some evidence that 'like things' show up close to one another in this space. The diphthongs are close to vowels; and they also share the property of not being steady-state with the plosives; the voiced plosives are closer to the vowels; and the liquids and nasals are also close to one another. Jakobson thought that the liquids and the nasals functioned as a natural class, and there is further evidence for this supplied in Anderson & Ewen with respect to the Dutch diminutive suffix selection (1987, p. 153). The Z direction is essentially inversely proportional to the signal-to-noise ratio, considering the formant peaks as the signal and the friction as the noise.

APPENDIX VI

Ordinal Vowels and Vectors

The vowels can be rendered as bitstrings representing distinctive features, or as vectors. As a 0th order approximation we can represent the vowels shown as bitstrings. We can add other vowels (for example æ) as in fractional or discrete coding simply by representing it as (1,1/2,0) which can be thought of as a vector. The approximate articulatory degrees of freedom (DOF) can be represented directly. From this we can easily extrapolate also to the others' schemes such as those of Trager & Bloch, or Chomsky, as can be seen in Hubey (1994). This simple discrete/binary coding can be thought of as a prelude to the vector spaces developed in the body of this article.

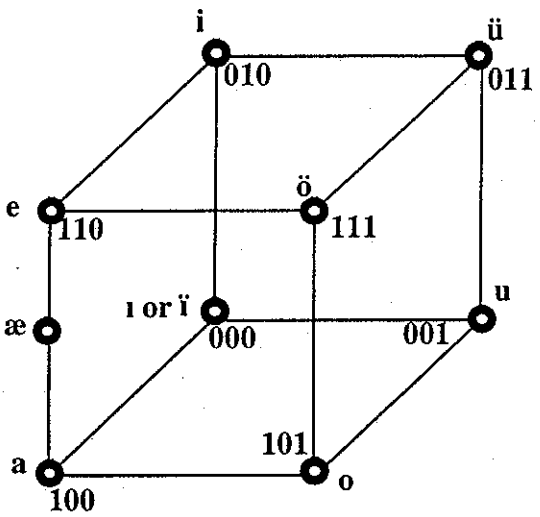

Fig. VI.1. The binary vowel space (Hubey, 1994).