The Failure of Phonetic Discreteness and Auditory Identifiability: Response to A. Manaster-Ramer

Robert Port*
Department of Linguistics,
Department of Computer Science
Program in Cognitive Science
Indiana University, Bloomington, Indiana 47405

June 26, 1996

Abstract

The phenomenon of incomplete neutralization and the subtlety of this incompleteness reveals vividly that speech sounds do not fall into discretely distinct phonetic types, and also that auditory impressions cannot be relied upon to identify them. AMR appreciates that these failures pose a serious problem for phonology. The reason is that the segmental units of standard phonology inherit the properties of discreteness and auditory identifiability from universal phonetics. So if phonetics is not always discrete, and practical identification is inherently unreliable, then phonology must restructure itself from the ground up. Phonology would then have no way to account for its discrete phonological objects (whether phonemes, rules, syllable trees, metrical grids, rule rankings, allophonic rules, etc). This seems to be the theoretical bad dream that concerns AMR. I will argue that all

*Thanks to Stuart Davis, Ken de Jong, Mike Gasser, Diane Kewley-Port and Charles Watson. This research was supported in part by the Office of Naval Research, N0003-1267.

this is true and that a explanation for the discreteness of phonology should be sought elsewhere, not in the hypothesized universal phonetic inventory. In fact, the explanation for phonological discreteness must be sought in the same place it is sought in other sciences (eg, in astronomy, meteorology, economics, nuclear physics, etc) – in the dynamically created structures often exhibited by systems with many degrees of freedom and a constant flow of energy.

1 Introduction

Are phonetic units discretely different from each other? Can careful listeners identify speech sounds reliably? These are the primary issues raised by Prof. Manaster-Ramer (AMR) in his letter to the editor. The author is concerned about a phenomenon reported a number of times in the phonetics literature 6-10 years ago. He is alarmed that these results may undermine the methodological and theoretical basis of phonology. In this essay, I will first review the phenomenon itself and agree with AMR that 'incomplete neutralization' is just a slight variant of many other familiar phenomena from experimental phonetics and phonology, and that it is troubling. Many widely attested phenomena offer strong evidence that the set of sounds of speech, that is, the universal inventory of phonetic categories, cannot be assumed to fall into discretely distinct types that can be reliably differentiated by a first language learner and by a trained linguist or phonetician. Of course, often speech sounds are obviously different (eg, the initial stops in *bad* vs. *pad* or the vowels in *bad* vs. *bed*). But can they *always* be assumed to be discretely different from each other and sufficiently distinct that a careful listener could hear the difference? Traditional linguistic phonetics says yes, but I think the evidence makes this very unlikely. That is, if you ask regarding two short sound bites "Are these two instances of the same sound or different ones?," one simply cannot assume that there will always be a correct or consistent answer. But if this is so, then one cannot use phonetics to explain the discreteness of *bad* and *pad* either. The evident discreteness of phonology will need to be accounted for in some other way.

AMR is quite correct to be concerned about these issues: *Failure of Phonetic Discreteness* and the *Failure of Auditory Identifiability* – about the fact that neither a linguist nor a native speaker can necessarily identify discretely and confidently which phonetic elements they observe in a sample of speech. This failure raises distressing questions about how the discipline of linguistics should obtain data regarding the phonological systems of human language. The phenomenon of incomplete neutralization *is* a worrisome chink in the dam that holds up all of current symbol-based phonological theory – just as AMR fears.

Many phonologists prefer to act as though one could assume confidently that the process of careful listening followed by phonetic transcription (at least when performed by a professional) 'makes available' to the phonologist a representation of *all possible linguistically relevant aspects of a sample of speech.* From this discrete symbolic description, the phonologist constructs various data structures, from phonological segments to syllabic trees to metrical grids. Without this discreteness and the presumed positive identifiability of these units, phonology would appear to have no place to begin to work. How could phonology account for symbolic structures like language-specific consonant and vowel types, allophones, stress markings, etc.? Where could these complex cognitive objects possibly come from?

Chomsky and Halle (C-H) 1968 offered an answer. They essentially proposed that the universal phonetic space is a discrete alphabet. There is some list of features that is the maximum set of linguistically relevant sound types in languages of the world. All members of our species are born with the ability to reliably and almost effortlessly identify these perceptual objects. These static features serve as units for language learning children and constrain the possible speech sounds of language such that they take the form of a two dimensional matrix of feature values versus discrete time. (Although C-H asserted that the features might have continuous values rather than discrete ones, linguists nevertheless invariably treat them as though they had discrete, and indeed binary values.) Thus, in effect, the data structures of phonological descriptions *inherit their discreteness from the phonetic atoms* of which they are constructed. Thus a basic discreteness is built into linguistic ontology by Chomsky and Halle. Whatever the physical world might be like, it is assumed that as far as language is concerned, nothing other than discrete abstract phonetic objects need to be the concern of linguists. Indeed it seems to be closely related to the Cartesian view that the human mind is quite a different sort of entity from the physical world.

For phoneticians, on the other hand, the failure of phonetic discreteness is in fact rather familiar. Phoneticians have long observed that phonetic objects are discretely different only in ideal circumstances (Lisker-Abramson, Keating, Lindblom, Klatt?, etc). However, although AMR is concerned about the partial neutralization problem and suggests that great amounts of experimental work might have to precede any linguistic analysis, he does not seem to put his finger on the magnitude of the theoretical threat. In this essay, I will try to clarify what is at issue and why. Finally I will suggest that a longterm theoretical solution to the problem of phonetic discreteness may actually be at hand.

2 What is 'incomplete neutralization'?

The phenomenon that concerns AMR can be illustrated with the German stems *Bund-* "brotherhood"and *bunt-* "colorful". If the stems have a suffix, as in *Bunde* and *bunte*, then they are pronounced [bUndə] and [bUntə]. The orthographic difference between the T and D is clearly appropriate, and illustrates one of the most striking and profound properties of human language: the use of discretely distinct sound types for 'spelling' lexical items. These phonological types, the D and T, recur in many different vocabulary items and often may appear in a range of syllable positions. However, when the /d/ or /t/ occurs as the last segment in a syllable (that is, for these stems when there is no suffix), the contrast is often said to be 'neutralized' since both words are pronounced as approximately [bUnt] with a final [t]. The D/T distinction appears to be lost since both pronunciations (for these words and for many other pairs of voiced and voiceless obstruents in Standard German) merge into a phonetic [t] (or the cognate voiceless stop, fricative or affricate). The traditional linguistic description of this phenomenon is to say that a 'neutralization rule' has applied that turned the [+voice] feature of the D into the [−voice] feature of the T.

The experimental surprise is that if you do either a production experiment or a perception experiment using a reasonable number of tokens of each minimal pair, you find that the two sets of [t]-like sounds are consistently somewhat different (Port and Mitleb, 1981 and replicated in Port and Crawford, 1989 and elsewhere). That is, production data show small but statistically significant differences in many measurable properties between the two underlying stops. The differences are typically in the direction one might expect from examination of obviously contrastive pairs like *Bunde-*

bunte. Thus the [bUnt] that is related to /t/ has a slightly shorter preceding vowel and slightly longer closure than the [bUnt] related to lexical /d/. Also, there is weaker pulsing during the stop closure and a somewhat stronger release burst on the stops – both of which imply that speakers tend to pull their vocal folds somewhat wider apart for the underlying voiceless stops than for the voiced ones. Of course much overlap remains between the two voicing sets for any of these measurements. But the word pair populations are clearly only slightly different, suggesting that the 'underlying' voicing feature is still biasing the phonetic detail of the stops despite the fact that most of the difference between the voicing classes has been wiped away when the stop or fricative occurs in this syllabic position. This result has also been reported by several other investigators (as mentioned by AMR and the earlier papers).

But can listeners ever make perceptual use of this difference? We found in the 1989 paper that if one records a number of productions of this pair and you ask native German listeners to identify them (in a forced choice between pairs like *bunt* and *Bund*), you find that - rather than 50% correct identification (indicating mere guessing due to the supposed neutralization) – one obtains roughly 60-80% correct identification of the word pair (depending, of course, on many linguistic and channel variables that affect the rate of correct identifications). But if subjects were instead identifying a typical minimal pair, such as *Bunde* vs. *bunte*, in these experimental conditions, they would probably identify them at around 99% correct. So the first thing that is puzzling about these data is that the traditional view about phonetics predicts performance close to either chance (= 50% in this 2-alternative case) or close to 100%. After all, two sounds are either exactly the same or else they must differ by at least one 'phonetic quantum'. And one should expect that any difference will be at least reliably detectible. If such a difference were not reliably detectible, then how could children learn their native language? And how could linguists do phonetic transcriptions that capture "all and only" the linguistically controllable phonetic differences?

Adding further strangeness to this case of 'moderately accurate' performance is, as AMR notes, that many well-known phoneticians and phonologists (who should be expected to be at least as good phonetic discrimination than ordinary lay speaker/hearers) have asserted that the German voicing contrast is 'neutralized' to some kind of stop that is the same for both *Bund* and *bunt*. Not only did the great German phonetician, E. Sievers consider the neutralization to be a complete neutralization, but so did L. Bloomfield, M. Halle and many others. The D simply turns into a T. So AMR is understandably puzzled at the paradox: "How could this difference be fairly easy for native speakers but seemingly be completely impossible for (or perhaps negligently overlooked by) highly respected linguists and phoneticians?" And the problem is compounded by claims of similar phenomena in syllable-final devoicing in Russian, Polish, Catalan, etc.

In fact, for me personally, one reason why the whole phenomenon seemed quite plausible is that soon after serendipitously discovering this phenomenon with Fares Mitleb, working on his 1980 dissertation, I noticed a very similar effect within my own speech! For me as in most dialects of American English, a very similar partial neutralization is found for the flapped medial /t/ and /d/. Thus, for me, pairs like *budding-butting*, *biddy-bitty* and *ladder-latter* all seem to sound the same (although for some Americans the last pair may exhibit a more noticeable difference due to difference in the pronunciation of the [æ]). Indeed in teaching introductory phonetics classes I had taught students for years that these pairs were neutralized! "Say them", I would say, "You can hear that they sound the same despite the spelling difference. This is neutralization," and the students would nod in agreement. So although *bud* and *butt* are obviously constrastive, *budding* and *butting* seem at first to be indistinguishable. However, Fox and Terbeek (1977) showed that many Americans produce these two 'flaps' with somewhat different timing (the preceding vowel tends to be shorter before the underlying /t/). I have found that there are also statistical differences in the stop closure duration and in the visible voicing during the stop closure (as can be seen on sound spectrograms) (Port, 1981). With isolated words read aloud from a list, it is fairly easy to demonstrate in the classroom that American listeners can guess with much better than chance accuracy whether *budding* or *butting* was intended. Several unpublished student projects over the years have demonstrated 60-70% correct identification for such 2-alternative foced-choice pairs in American English.

So the English flapping situation and German syllable-final devoicing are very similar. The primary difference is that in English both the /t/ and /d/ are modified (in the post-stress, intervocalic context) to a third sound, the apical flap, whereas in German the voiced obstruents seem to merge directly into the voiceless ones. Both cases are effective enough as neutralizations that native speakers (like phoneticians) do not immediately notice the difference.

And in both cases, there is clear morphological support for the underlying contrast - that is, *Bund-bunt* and *budding- butting* are only homophonous in certain morphological contexts.

In the next section I take a closer look at the traditional view of phonetics as presented by Chomsky and Halle, 1968 in order to show how this problem arises.

3 The Standard Theory of Linguistic Phonetics

AMR takes the traditional linguistic approach to speech perception very seriously. This theory of linguistic phonetics is essentially the one presented in Chomsky and Halle in their 1968 *Sound Pattern of English* (SPE, especially pp. 293-301) - and is derived from earlier feature theories (eg, Jakobson-Fant-Halle, 52; Pike, 19nn, Hockett, 19nn). Although many specifics of phonology have changed over the years, it is difficult to see much change in the treatment of phonetics from within generative phonology since 1968. Only recently in 'gestural phonology' (eg, Browman and Goldstein, 1995) and in the movement toward 'laboratory phonology' has fundamental change occurred. I will try to summarize this standard theory of phonetics as viewed by many linguists.

The whole issue derives from the Chomskyan distinction between *Competence* and *Performance*. When it comes down to mechanisms, this is normally interpreted to differentiate between Competence as the domain of *discrete variables* (that is, symbols and symbol structures) as they are configured and reconfigured in *discrete time*. 'Processing time' involves discrete jumps between system states when a rule is executed. (The structure of events in real time associated with the pronunciation of words is also discrete but is encoded as the ordering of static objects like segments, words and other syntactic units.) In opposition to this is Performance, the domain of *continuous variables* evolving in *real (continuous) time*. The continuous variables include processes related to motor control, audition and speech perception. So Performance is essentially the physical, while Competence is cognitive or mental and is assumed to work on principles similar to those of logic, mathematical proof and digital computation (See Haugeland, 1988 AIVI and van

Gelder and Port, 1995). Within linguistics, it is an article of faith that language (and probably everything else that is mental) will be best understood in terms of discrete time and discrete symbols.

But the competence-performance distinction (closely related to the Cognitive vs. Physical and the Mind vs. Body distinctions) creates problems at both output and input. The first is how can discrete, static symbols independent of real time (in the mind) control realtime continuous performance (in the body)? Presumably the mind controls the body that produces the speech gestures. But, as pointed out by Fowler et al, 198x and Turvey on Bernstein, any model of this process must be quite implausible. The problem is that temporal specifications must now be laid onto all those timeless symbols at output time and then performed somehow. It is very difficult to imagine how durational specifications could actually be performed that is not full of arbitrariness. In trying to do it, one is forced to continue discovering (or is it inventing?) new static states, thus causing the problem blow up (see, eg, Port, Cummins McAuley, 1995). Then some executive system from within performance must assume responsibility to assure that each minisegment type actually lasts the specified amount of time. The second problem, the input problem, is how can the auditory system become a phoneticizer and translate continuous-time auditory events into discrete static symbols? Mechanisms capable of this can be easily constructed, but how could they be designed so as to exploit all the many kinds of subtle temporal information that human speakers and hearers employ? Here too, new intermediate and context-sensitive states tend to proliferate as soon as one looks closely at any data.

Of course, Chomsky and Halle did not need to address these performance problems in order to do phonology. They only needed to have some description of speech articulation and speech perception that was sufficient to characterize the sound systems of human languages. They assumed, naturally, that this would take the form of a list of symbols, an inventory. So they asserted the existence of an *interface alphabet*, the list of 'the phonetic capabilities of man' as they called it in their ringing (but now quaintly old-fashioned) turn of phrase. These minimal phonetic objects are atomic symbols as far as Competence is concerned, even though within Performance it is assumed they have both articulatory and auditory aspects involving continuous variables and continuous time – very difficult problems that were left to the phoneticians to deal with.

Thus, on the motor side, these discrete phonetic objects can be thought of as providing a universal alphabet of control configurations – all that is necessary for speech production specification and all that could in principle be controlled by the grammar of a language. Nothing beyond these units (that is, requiring more articulatory or acoustic detail) is supposed to be controllable – at least, not by the grammar of a language (though obviously people can mimic each other in nonlinguistic ways). On the perception side, it is assumed that audition comes with a 'phoneticizer' that exhibits 'categorical perception' and translates continuous acoustic events into discrete symbolic descriptions. It is these two devices, the input device and the output device, that are responsible for phonetic discreteness. And it is their performance that is thrown into question by incomplete neutralization (as well as by many other data, of course).

The traditional linguistic theory of phonetics and speech perception thus asserts that:

1. humans can hear the sounds of human speech (only) in terms of a set of discrete sound categories usually called 'phones',
2. each phone is a simultaneous combination of phonetic features,
3. each phonetic feature is a static, atomic object with a simple, unitary articulatory and auditory specification and no internal temporal structure,
4. there is some closed set of such features for human language,
5. individual languages employ subsets of the universal set for their lexico-phonological systems.
6. children learning their language employ these units to organize the perception of speech and the phonological grammar of the language

It follows that two phones may be either identical or distinct. If they have the same phonetic features, then they should be phonetically identical – that is, as far as linguistic control is concerned. (Wouldn't this imply that no human should be able to distinguish them perceptually?) If they are different (that is, have distinct phonetic features), then, at the very least, language learning children and native speakers should be able to differentiate them easily and reliably. Why? Because if children could NOT be counted on to make the appropriate distinctions for any language, then how could various languages be reliably and accurately acquired? That is, the universal stability of the phonetic units is what provides the explanation for the universal stability of language acquisition.

Of course, it is well known that many sound units in a non-native language may be quite difficult to acquire when learned by adults. So one must suppose that adults may fail to distinguish the novel sounds of languages due to having somehow lost much of their innate ability to recognize the sound distinctions of human speech (Werker et al, Yamada et al...). Apparently for most lay speaker-hearers, only the phonetic categories used in their native language are normally useable for speech perception as adults.

So what about phoneticians and linguists? How do they evade this phonetic atrophy? Presumably the atrophy effect may be reduced by phonetic training. The International Phonetic Association alphabet (IPA) and Chapter 7 of SPE are two examples of scientific attempts to organize and list the full set of controllable aspects of speech perception and production. These are supposed to be all that is potentially under linguistic control - whether for contrasting words or for simply controlling the motor system. Linguists and phoneticians attempt to retain the distinctiveness of these features so as to produce and perceive them. It would seem that a strict version of the standard phonetic theory should predict that humans can never hear *more* detail than what is provided by the 'phoneticizer' (though little is normally made of this). An implication of the notion of a universal phonetic alphabet is that we may hope that professionals linguists and phoneticians should be able to *approach* the perceptual and motor skills of native speakers of all languages.

The Chomsky-Halle model recognized that beyond Competence there is Performance, and acknowledged that the physical signal supporting speech perception is characterized by continuous change over time (eg, formant trajectories that result from an articulatory motion) and by continuously variable parameters (such as formant frequencies, intensities, lip positions, etc). However, C-H are very clear that *the only aspects of continuous speech events that could be relevant to linguistic competence are those differences that reflect distinct phonetic transcriptions.* The grammars of specific languages can only use some specific universal list of phonetic elements.

The fundamental reason for making this bold assumption is the one pointed out by Haugeland (1985, pp. 52-58): *symbolic theories simply must assume a set of positively identifiable symbols.* That is, the formal system itself – the grammar – must have symbolic objects that are discrete. They must be discrete in order to be infallibly recognizable. The symbols must also be stable over indefinitely long periods of time: if you put a symbol somewhere, it must still be there when the system comes back later to read it. Formal models intrinsically depend on these properties in order for their rules to function properly and for data structures to literally hold themselves together. In the execution of a computer program (one a familiar example of a formal system), these stabilities are assured due to the engineering of the chip. For cognition in a human brain, if it is to also be a genuine 'competence model', as Chomsky and Halle clearly intend it to be, then these properties must be assumed. Without discreteness, infallible recognition and indefinite time stability, symbolic or computational models simply will not work. Rules cannot executed if the system cannot be sure when it is looking at an A rather than a B.

So C-H had to propose that there is some phoneticizer that chops messy speech into useable symbols. It is assumed that only the output of the speech perception mechanism is available to any language, and this output must be constrained to provide only atomic and static phonetic feature symbols selected from the universal set.

This theory, the standard theory of linguistic phonetics, and has remained essentially unchanged in the phonological literature since the mid-1960s although it is probably fair to say that C-H primarily formalized ideas that were current from the 1930s on. The most fundamental problem with Competence is that within competence there is a sharp distinction between Symbols (as indefinitely time-stable states of the system) and the instantaneous Transitions between states. This idealization of processing within competence pretends time is nonexistent. Time exists neither in the Symbol nor in the Transition. But real time moves inexorably. And it can be looked at over much shorter (or much longer) time scales. Events that seem 'instantaneous' to our cognitive intuitions may look extremely long at the much faster time scale of neurons. The computational paradigm assumes that whatever might be happening at any shorter time scale (eg in neurons, or generally, in Performance) is perfectly irrelevant for what happens at the Competence time scale.

By idealizing Symbols and Transitions in this way, computational models of language make continued scientific progress on a theory of language very difficult. A theory of phonology (and of linguistics as a whole) that can be incorporated into modern cognitive science must begin with a far more sophisticated view of the relation between cognition and the physical aspects of the body and the physical world than merely the simple mapping of performance symbols onto competence symbols offered by an any interface alphabet. Insisting on a mere mapping relation between Competence and Performance makes it impossible to understand how language is situated in a nervous system. A practical approach should not assume that linguistics is the study of the symbolic and formal structures of languages, but rather it should view linguistic structures of all kinds, from phones to words to sentences, as events in time. Different kind of structures 'live' on a range of time scales (eg, sentences are longer than phonemes). Of course, we now have the technology to write words down on paper or put them in a computer file. We can scan these displays in both directions and look for patterns. But such a display is not available – at least not apriori – to human cognition. If such a spatial display of words and sentences exists cognitively, then accounting for how it could work is a scientific problem. However, to assume that such representations exist is theoretically reckless. As used by human speakers and as experienced by cognitive systems, *the true dimensional axis of speech and language is time, not space.*

It is the erroneous assumption that 'language is a formal symbolic system' that forces phonology to insist that linguistic phonetics must provide discrete universal objects. It is just that phonology needs something formal to manipulate. Of course, there are also a number of other arguments for discreteness that have appeared along the way. Although they are often put forth as relevant evidence by phonologists and phoneticians, none of these empirical arguments, in my opinion, have much relevance to the central issue. Nevertheless, it is worth considering these performance-related arguments that seem to be related to the assumption of discrete phonemes.

4 Empirical Arguments for Phonetic Discreteness

There are three main arguments that seem to be proposed most often for why we should believe that the discreteness assumption is justified. They are that (a) the *limited perceptual resolution* of humans forces a limit on the number of distinctions possible along any dimension and thus supports discrete categories, (b) the phenomenon of *categorical perception* is well attested and suggests that discrete categories with sharp perceptual boundaries are intrinsic to speech perception, and (c) evidence for the *quantal theory of speech* further shows that discreteness is natural for many phonetic categories. I think that none of these provides any compelling evidence that the phonetic space is a universal and invariably discrete inventory.

Limited Perceptual Resolution Argument. To many phonologists the discreteness of speech perception seems almost to require no defense beyond common sense. After all, so the reasoning goes, one can only distinguish sound differences with a certain level of detail. So it seems that only a certain number of, say, vowel distinctions should be possible due to finite limits on resolution on auditory sensation. But is this reasoning sound? Limits on resolution do not necessarily yield discrete levels. In the late 19th century similar reasoning led early psychologists like Titchener and Wundt to similar conclusions about simple stimulus scales such as pitch, color, and so forth (see Boring 1943 'Sensation-Perptn').

For example, Titchener and others viewed pitch perception as reflecting a sound unit called the 'just noticeable difference' (JND). The idea was that the frequency scale for pure tones, for example, must be divided into discrete steps (just like pixels on a computer screen but in one dimension). Thus, if two tones are presented to a subject one at a time that are close enough along the scale to fall within the same JND, then it was predicted that they should be reported as the same, but if they lie in different JND regions, then they should be reported as different. This is similar to the kind of reasoning that Chomsky and Halle employ in concluding that vowels must have a fixed (small) number of discrete height values, and similar to AMR when he supposes that either *Bund* and *bunt* are perceptually the same or they are different.

But psychology abandoned the JND view of pitch resolution long ago. The primary reason is the ubiquity of noise internal to the perceptual system. Thus it is clear that listeners, even for a very simple discrimination task, do not always give the same response when the discrimination is difficult. So if you ask them 10 times about the same pair of stimuli, you will often get some 'sames' and some 'differents'. If one begins with an impossibly small difference and increases the stimulus difference toward easier discriminations and plots the probability of saying 'different' (from 0 to 1) against the stimulus continuum, the data will, essentially always, sketch out an S-shaped curve – with higher probability of saying 'different' correlating, of course, with larger changes in the stimulus.

In fact, if you don't get a smooth curve, then you have either not been sampling closely enough along the stimulus continuum or else have not looked at enough tokens (either within or between listeners). Can one locate the boundaries between the hypothesized discrete sensory categories? For a small enough difference, moving the difference along the continuum, one should (on the discrete category view) find discrimination flat spots (where the sounds are within the same sensory class) alternating with discrimination bumps (across a boundary between the sensory steps). But they are not found. These days, when a psychophysicist speaks of a 'just noticeable difference', it is interpreted to mean enough difference that subjects have, say, a 75% chance of detecting the difference. The S-shaped psychometric function rules out any nonarbitrary steps along stimulus continua. Of course, this is just as true of speech stimuli as for anything else. So the fact that there are sensory limits relevant to distinguishing phonetic categories from each other in no way justifies a claim that there is a discrete set that can be reliably identified.

The notion of 'reliable identification' runs into another difficulty as soon as probabilistic judgment appears. It turns out that one must differentiate the *'sensory analysis'* aspect of the discrimination task (or the identification task) from the *'response decision'* aspect of the problem. Subjects may have a bias toward one response and against the other. For example, if subjects are asked to make a discrimination that is difficult enough that there is some chance that either response might be correct, then other criteria will play a role in determine their response. These are usually called response biases. For example, if the payoffs and penalties of the situation are such that making a 'False Alarm' (calling them 'same' when they are not) costs more than the reward for a 'Hit' (calling them 'same' when they really are), then observers will be tend to be conservative about responding 'same'. These decisions will be affected by a number of features, including the subject's estimate of the apriori probabilities of one state of affairs vs. the other (thus, for example, if subjects expect to see more /t/s than /d/s, they they will adjust their response criterion to make sure they respond /t/ more often). So, the actual response of subjects (and therefore their actual percent correct in a discrimination or identification task) depends *only in part* on the results of their perceptual analysis of the physical stimulus.

This problem is actually fairly easily solved: The Theory of Signal Detection (TSD) (see Swets and Green, etc) has demonstrated statistical methods to correct for response bias (by taking into account the proportion of hits to false alarms) and has suggested experimental procedures that permit a bias free measure of the distinctiveness of the two classes can be obtained. (See Port and Crawford, 1989, for a simple application of signal detection theory to our German voicing contrast results).

All of these same features apply to speech. The best general assumption about the perceptual mechanism is perhaps that it produces a probability judgment about a stimulus with respect to several possible categories. Which response a listener (including even a phonetician or linguist) actually gives may reflect a variety of factors that have nothing to do with their perceptual similarity.

The Categorical Perception Argument. It has been known since the late 50's that if you vary speech stimuli along complex continua between phonetic classes, subjects' perception will jump rather discretely (Liberman et al, 1968; Harnad, 1991; Repp, 1991). Is this good evidence of a discrete 'phoneticizer' in speech perception? No, it is not, for one simple reason. Standard linguistic phonetics requires that ALL phonetic contrasts be discrete while the categorical perception effect has been known from the earliest days to occur only for certain subclasses of sound contrasts (eg, for place of articulation and voicing) but not for vowels. Yet discrete reliable categorization of vowels is every bit as important for phonology as for consonantal features.

Of course, categorical perception is a much more complex problem than distinguishing pure tones differing in frequency. Speech stimuli have enormous complexity and richness, but on the other hand, they receive a huge amount of practice too. It is pretty clear that when we present listeners with very complex stimuli, only certain aspects of the stimulus tend to be heard accurately (Watson). Most details cannot be noticed. On the other hand, it is known that if you give subjects a great deal of practice on just a single speech stimulus, listener's can respond in ways that reveal that their auditory resolution approaches the limits that observed for simple tones (DKP-Watson).

So categorical perception is still problematic and not understood, but it is clear that it does not provide any justification for assuming that all speech sounds are discretely and reliably perceivable.

The Quantal Theory of Speech Production Argument. K. N. Stevens (1972, 1989) demonstrated that the acoustic response of the human vocal tract behaves highly nonlinearly for certain changes in articulation. The consequence of these nonlinearities is that for speech sounds at certain locations along articulatory continua, any variation in articulatory accuracy will have minimal consequences on acoustics. Other related arguments from various phoneticians (eg, Lindblom, 1989) have developed analogous arguments about the perceptual distinctiveness of, for example, vowels. The 'quantal properties' of speech suggest that certain places of articulation, certain manners of articulation and certain vowels are relatively insensitive to articulatory errors. Stevens argues that somehow this justifies or rationalizes the assumption of discrete phonetics. But his evidence really only supports the claim that, given some reasonable assumptions about articulatory and auditory preferences, certain speech sounds may be more 'attractive' than others. That is, because of these nonlinearities certain sounds may be more efficient choices for languages to employ than others. This comes nowhere close to providing empirical support for the traditional phonetic claim that there is a discrete, reliably identifiable, innate and universal sound inventory.

A more plausible implication of Stevens' arguments is that it explains why certain particular sounds appear in language after language: one thinks of [s,t,d,D,n,a,i,ə], etc. These sounds may be both articulatorily and auditorily advantageous.

In short, none of the empirical, performance-based arguments are even directly relevant to the assumption of a universal, reliably identifiable phonetic alphabet. The fundamental rationale for such an alphabet is really only the theoretical one that the study of competence phonology cannot begin without an alphabet.

Phonologists would like to believe that experimental research justifies this assumption, but it does not. In fact, human speech perception is unreliable and nondiscrete. Speech production is the same. Certainly it contains enough noise that speakers' productions of the same linguistic identities always span some range if careful measurements are made, and speakers clearly have control over continuous variables that permit sounds to exhibit distributions that overlap to any degree – from statistical identity to being obviously very different. But there is no reason to believe that all these difficulties are swept away by some 'phoneticizer'.

But one might say that even if all this is true, there is still no reason to suppose that anything important has been lost by the assumption of segmental transcription as the basis for phonology. So what is the problem?

5 Consequences for Phonology: What is Lost?

I am suggesting a rather sceptical view of the process of data collection that is frequently employed in phonology. It seems that this process is roughly that linguists (and sometimes phonetic specialists) produce phonetic transcriptions that represent in symbolic form all the 'linguistically relevant aspects' of speech production and perception. Then phonologists employ these transcriptions as their 'primary data' for generating models of the data structures and language-specific processing of the language. Although these days there are many 'experimental phonologists' or 'laboratory phonologists' who combine careful data collection on speech sound and on motor control (see LabPhon 93-96, Beckman-Kingston), there remain many phonologists who feel that their work addresses phonological questions for which impressionistic phonetic transcriptions will serve as completely appropriate data. But in my view all phonologists should be at least a little worried about this assumption.

After all, incomplete neutralization is just one of a vast variety of experimental phenomena collected over roughly 50 years of intense experimental observations of human speech. These result show, for example, that

- there are many kinds of subtle context effects that appear in the study of speech articulation or speech acoustics (to get a smattering, see, eg, Cole-Zue on SG reading; Liberman et al 1968),

- many of these context effects are manifested in the time domain (eg, Lisker-Abramson, Lehiste, 1968, Klatt, 1976, Port, 1981), and
- many of these effects are specific to particular languages, rather than being phonetic 'universals' (Port-Alani-Maeda, etc).

It is these subtleties that have made automatic speech recognition so challenging. Furthermore,

- essentially any of these subtle variables can, under appropriate conditions, be employed by listeners in speech perception (too many references to cite).

So, I suspect that there are a great many phonologically important aspects of languages that are completely missed by reliance on the traditional symbolic phonetic transcription of speech.

Let me mention briefly here two specific examples of phonological phenomena in areas I am familiar with that slip through 'the segmental grid' - problems that seem to me to be ignored or dealt with awkwardly in phonology due to a primary reliance on discrete phonetic transcriptions.

The Germanic Postvocalic [Voice] Contrast. In English when the voicing contrast occurs at the end of a syllable, as in the pair *fuzz* and *fuss*, the difference in voicing is manifested as a change in the ratio of the duration of the vocalic part of the syllable to the duration of the final consonantal portion of the syllable (eg, Port,1981; Port-Dalby, 1989, based on earlier results by Lisker, 1955, etc). This durational ratio also helps to characterize the contrast between, say, *bids-bits, camber-camper, Libby-lippy, Bangor-banker, large-larch, ruby-rupee*, etc. A similar contrast in the vowel/consonant ratio for distinguishing voicing pairs is also found in many other Germanic languages (at least Standard German, Bavarian, Swedish and Icelandic). But if you just do a segmental transcription to represent the data, then the durational difference between the duration of the stop and fricative closures (eg, between [t] and [d] or [s] and [z]) seems uninteresting (because it affects only 'phonetic implementation', not the phonology) and the effect on the preceding vowel and any voiced consonant (eg, *lunge-lunch*) is just another instance of a context-dependent phonological rule, of which there are thousands of examples. The compensatory or inverse durational relationship is completely obscured – due entirely to the dependence on segmental transcription.

So here is an important language-specific phenomenon, with many variants across the Germanic family (including the very "near neutralization" phenomenon that stimulated AMRs essay). This shortening followed by lengthening (looking at voiceless with respect to voiced) could be viewed as a brief warping of speaking rate (Port-Cummins, 1989). But however this ratio effect should be described, it is clearly Phonology - part of the 'grammar' of English, German, Icelandic, etc. Most other languages do not show evidence of manipulation of these ratios. Study of the temporal structure of English and German (and other languages from this family) should reveal something fundamental about how linguistic information regarding classes of lexical items can be encoded by languages. Apparently abstract temporal ratios, or "intra-syllabic rate-warping" are one practical parameter 'invented" by this family. This seems to be, on the face of it, a fascinating phonological problem. But transcription of these languages as exhibiting a mere segmental contrast in [$\pm voice$] with the treatment of the vowel duration as a feature change and the consonant duration as outside phonology altogether makes it impossible for most phonologists to see the abstract temporal relationships that underlie the voicing contrast in this family of languages.

English 'Meter'. A second domain where the description of speech in symbolic terms may render important phonological structures invisible is in the problem of meter. English phrases often seem to have a global timing structure over a scale of a second or more. Phrases like *Mississippi legislators* seem to have an alternating pattern of stresses – whether 4 of them or only 2 (see Hayes, 1995). It has been proposed many times (Jones, Abercrombie, Martin-72) that music-like rhythm, definable in terms of relative duration, may underlie such pronunciations. Of course, actual temporal studies often find messy and unclear results (Lehiste, van Santen-96, etc). Thus, phonologists will often address the problem of meter with a discrete time scale, using one 'time step' for every syllable (eg, Hayes, 1995). But will discrete time prove sufficient for an understanding of English metrics? This seems unlikely.

In recent experiments in my lab, we have explored the temporal aspects of English metrics (Port-Cummins, 19nn, Cummins-Port ICSLP). We first made the hypothesis that some sort of real-time oscillatory system might underly the metrical aspects of speech timing. If this is so, then we should be able to interfere with such an oscillatory system by encouraging coupling with another oscillatory pattern (inspired by Kelso, eg, 1995; Treffner-Turvey, etc). To illustrate 'coupling', imagine a parent pushing their child on a swing. The parent will couple their body motions to the rate and amplitude of the oscillating child-swing system. If the swing length changed or if weights were put on the seat, then the parent will adapt to the change in frequency – that is they will remain coupled to the child-swing oscillator. This state of coupling provides the most efficient way to use their body to keep the swing going. Coupling is found both within our bodies (eg, between your legs most of the time, and between the neck muscles and the arm muscles while repeatedly brushing your hair) and between our own body and that of others (eg, communal singing or marching).

So, for example, imagine tapping your finger on the table from a comfortable position. If you are asked to tap the index finger on your left hand at some rate, you could do so fast or slow, at any rate over a broad range. But if you are asked to also oscillate your right index finger at some rate, then it turns out that you will only be able to perform both tasks stably at a small set of rates. They will be rates such that there is a very simple ratio, such as 1:1, 1:2 or 1:3 (or, with practice, 2:3), between the two fingers. Apparently the two fingers cannot avoid coupling with each other. They 'want' to keep certain simple temporal and phase relationships. (As an aside, notice that here is a case where the dynamics of interacting continuous systems produces a structure that is quite discrete and also stable over time – two of the useful properties exhibited by the symbols of formal systems! My hunch is that 'phonemes' and such will turn out to get their discreteness and stability in a related manner.)

We reasoned that evidence of coupling between a driving signal (a stimulus) and human behavior can be taken as evidence that the behavioral system incorporates an oscillator - rather like the arms are controlled by oscillators. So in our experiments we asked speakers to listen to a metronome signal (as a driver) and to repeat a simple phrase. Thus they might say 'Talk of the game' once for each tick of a metronome. This phrase has two metrical feet: 'talk of the', and 'game'. Not only do speakers align 'talk' with the metronome pulse (just as we instructed them to do), but the onset of 'game' has a very strong tendency to fall at certain phase angles rather than others, especially 1/2 (but also at 1/3 or 2/3) of the 'talk-talk' cycle. (The reader is encouraged to try repeating this phrase. You will probably find that the perceptual beat of 'game' is located half way between the phrase onsets.) So by means of this task - repeating a phrase to a metronome - we demonstrate a strong tendency for the 'foot oscillator' to entrain itself with the 'phrase oscillator'. We take the evidence of the ease with which speakers can couple their speech to a metronome and the tendency for feet to couple with the longer phrase unit to suggest that 'nested oscillators' running in continuous time underlie the metrical structure of, at least, English – whether or not there happens to be a metronome to couple with. If there is one, eg, if you are also tapping your finger or your fist on the table, marching with your feet, jogging, talking, chewing gum, or whatever, then speech will tend to couple with it. Otherwise, I suspect that global speech timing is controlled by a free, rate-varying oscillator-like metrical system.

This research program is only at its beginning, but a little observation of speech shows that in most cultures one finds a variety of special genres of speech that are rhythmic to varying extents. Think of songs, poetry, and chant. Think of a Bible-thumping preacher, an auctioneer, a TV pitchman or a rap singer. Though of course not all speech is equally rhythmic, it seems likely that the temporal constraints of rhythmic speech may offer a useful framework for understanding the less rhythmic aspects, rather than working only the other way around.

Of course, none of this implies that the work of, eg, Hayes using discrete time as the basis for meter, is not worthwhile. But it seems to me that a real understanding of meter will come only when Hayes' insights can be understood or reinterpreted in terms of a dynamical model of meter for English. When that is attempted, I suspect that some current issues will lose their interest (eg, 'rhythm-rule' phenomena will be easily understood) while other phenomena will find insightful new interpretations from the dynamical perspective.

I have argued that reliably identifiable and discrete phonetics is a necessary assumption for a formal or competence model of phonology and linguistics, but that no such reliable state-based speech perception is possible no matter how many years of training you have. Instead, we propose that descriptions that incorporate time directly by using the mathematics of dynamical systems will help us avoid error.

But there is still one essential task left to do in this essay. This is to offer a specific account of how the phoneticians could miss something that the native speakers could hear fairly easily.

6 Can phoneticians fail where native speakers succeed?

How could it be that phoneticians differ so sharply from lay native speakers in their ability to perceive this distinction? Given all that has been described so far about probabilistic perceptual outputs and about how response choices are affected both by stimulus analysis as well as other criteria, the answer is quite simple. They performed completely different tasks. Phoneticians and linguists are professionally concerned with establishing and using consistently a set of phonetic units that are plausible candidates for 'the universal phonetic space'. They would be understandably reluctant to claim to have discovered some categorical difference that they can only differentiate with, say 70% accuracy. That is, there are many criteria aside from the mere perceptibility of a difference that are relevant to their decisions about a phonetic transcription. We might say that they have a professionally motivated response bias. Only differences that are large enough and reliable enough are of interest. Thus it is important for them to ignore mere token-to-token variation, speaker idiosyncrasies, minor dialect differences, and so forth. Very likely these phoneticians and phonologists agree with C-H that any distinction that is to be represented in the 'universal phonetic alphabet' must be large enough that it might be useful (in some imaginable situation) within a language to distinguish words with fair reliability. From the perspective of these criteria, there is no question that the two [t]s in *Bund* and *bunt* should *not* be differentiated. They are indeed too similar to be of contrastive use.

On the other hand, in our discrimination experiments with native speaking Germans, the question we asked them was "Which word did the speaker say?" – rather than "Which letters would you use to describe these sounds?" So these listeners used all the acoustic evidence they could find to make their guess. There is no question in my mind that any of the phoneticians cited by AMR, even if not native speakers of German, could do as well as our German listeners at the word identity task. I bet that with a hundred training trials with feedback, almost any linguistics student could do well at this. I am not a fluent speaker of German but was able myself to perform as well at the task as our better German listeners. Like our subjects, I knew that there was an equal number of underlying /d/s and /t/s, so I just listened for *more* D-like vs. *less* D-like stimuli. The point is that choosing a phonetic transcription and identifying a word are two vastly different tasks with little in common except the stimuli themselves. It is little wonder that the phoneticians ignored these minute differences in assigning their transcription, and little wonder that the native listeners did use this information in identifying the words.

The view within traditional phonology that phonetic transcriptions can be reliably and discretely assigned is naive, it seems to me. What seems to be happening in the field is that the discipline of phonology is in the process of splitting between those who insist on doing careful experimental studies of speech and those who insist that impressionistic phonetics, typically obtained largely from secondary sources, is good enough.

7 Toward Phonological Morphogenesis

What I have done so far is to argue that there is no reason to assume discrete categorization in phonetics. The implication for linguistic theory is that this pulls the discreteness rug out from under phonology. If phonetics cannot explain the discreteness of phonology, how can we account for the discreteness we observe there? Languages obviously exhibit discretely different places of articulation, manners, vowel heights, etc in the units used to spell lexical entries. Where do these come from?

If one assumes that cognition is a continuous-time dynamical process, then a new set of theoretical tools become available. Morphological structures arise in many domains on many temporal and spatial scales, from the stars and spiral galaxies of astronomy, to species and self-organizing skin patterns in biology, to molecular valences and electron rings in physics. In some situations objects become structured as time-stable symmetrical patterns out of a flux of dissipating energy. The methods by which such self-organizing dynamics takes place are beginning to be understood in many domains (see, Thom, 1975; Murray, 19nn; Kelso 1995; Haken; Andrade?). There is no reason why linguistic structures should not be related to these other forms of physics (Port, et al, 1996 CLS). Such systems run in continuous time and follow dynamical laws and may produce many kinds of discrete objects. Instead of offering linguists a specification of parameters that some motor system is supposed to carry out in real time, the dynamical description will show just how it can be done.

8 Conclusions

This essay has covered a lot of ground. Probably too much. But I wanted to provide a good answer to AMRs worthy question.

First, the incomplete neutralization of German syllable-final devoicing is a real phenomenon. It has now been observed many times. The American English flapping situation is another very similar example. There are many others (eg, *prints-prince* near neutralization in American English (Fourakis-Port,19nn). I certainly do not insist that true neutralization cannot occur.

Secondly, AMR is correct to be troubled about incomplete neutralization. This effect, like so many other instances of nondiscreteness from phonetics, undermines a critical assumption about phonetics that is taken to be 'gospel' by many working phonologists. It seems to me (as it does for many phoneticians and 'lab phonologists') that, by trusting segmental transcriptions, they run the risk of building their work on sand. For example, such work tends to completely ignore everything about time except for serial order. And they tend to ignore or represent only clumsily all the various graded effects (like 'weakening' and so-called 'fast speech rules'). By ignoring these continuous effects and continuous time, phonologists run the risk of developing theories of the wrong phenomena.

Finally, as for phonetic transcription, despite my objection that discrete phonetic transcription is always unreliable, I will continue to teach the IPA alphabet to new linguistics students. It is useful for many purposes. But I also teach them not to trust it and not to imagine that their or anyone else's transcriptions provide reliable data on which to study many phenomena in language sound systems.

9 Bibliography

Abercrombie-text
Bloomfield-29
Boring, 1943
Browman-Goldstein
Chomsky-Halle, 1968
Cummins-ASA
Fourakis-Port

Fowler-Rubin-Remez-Turvey
Fox-Terbeek
Harnad-81
Haugeland, 1985
Hayes-95
Hockett - phonol
Jakobson-Fant-Halle52
Jones-bk
Keating
Kelso-95
Kewley-Port-Watson-95
Klatt-76-timing
Lab-Phon(Beckman,Kingston)
Lehist-book
Liberman
Lindblom
Lindblom-89-quant-theory
Lisker, 1955
Lisker-Abramson, 1970M
anaster-Ramer, letter
Martin-72
Mitleb-thesis
Murray-text
Pike - phonet
Port, 1981-lingc-timg
Port-Cummins-ICSLP
Port-Cummins-Gasser-CSL
Port-Dalby-89
Port-Crawford
Port-Cummins=Gasser
Port-Mitleb
Port-van G book
Repp, 1991
Sievers,E
Stevens-72-89 Quantal-theory
Swets-Green
Thom,1975-Semiophysics

Turvey - Bernstein
vn-Gelder-Port - Intro
Watson - review
Werker
Yamada et al -R/L
Zue-Cole