THE LACK OF INVARIANCE PROBLEM AND THE GOAL OF SPEECH PERCEPTION

Irene Appelbaum

Department of Philosophy
Washington University

ABSTRACT

The overall goal of speech perception research is to explain how spoken language is recognized and understood. In the current research framework it is usually assumed that the key to achieving this overall goal is to solve the lack of invariance problem. But nearly half a century of sustained effort in a variety of theoretical perspectives has failed to solve this problem. It is argued that this lack of progress in explaining speech perception is not, in the first instance, due to the failure of individual theories to solve the lack of invariance problem, but rather to the common background assumption that doing so is in fact the key to explaining speech perception.

1. THE CURRENT FRAMEWORK

The overall goal of speech perception research is to explain how spoken language is recognized and understood. In the current research framework it is usually assumed that the key to achieving this overall goal is to solve the *lack of invariance problem*. The lack of invariance problem arises in response to the widespread recognition that there is no simple mapping between units of phonetic structure and units of acoustic structure. A single phonetic segment is often realized by different acoustic signals and a single acoustic property may specify different consonants in different phonetic contexts. To rehearse familiar examples, the primary acoustic cue for the [d] in the syllable [di] is a rising second formant transition, while the [d] in the syllable [du] is signalled by a falling transition (Liberman et al. 1967). A single burst of noise at a frequency of 1440 Hz will be heard as a [p] in one phonetic context -- when followed by an [i] -- but as a [k] in another context -- when followed by an [a] (Cooper et al. 1952).

In the absence of obvious invariant acoustic properties with which to identify phonetic percepts, a central assumption of speech perception research has been that the goal of a speech perception theory should be to identify a set of invariant properties elsewhere in the speech chain. Thus, the *motor theory of speech perception* claims that phonetic segments are to be identified with invariant neural properties; the *ecological approach to speech perception* claims that they are to be identified with invariant articulatory structures; and the *theory of acoustic invariance* remains committed to the view that the invariants in speech perception are acoustic properties -- though of a different sort than those made salient in traditional speech experiments.

But nearly half a century of sustained effort in a variety of theoretical perspectives has failed to solve this problem. Indeed, not only has the problem not been solved, virtually no empirical candidates for solving the problem have been produced. One explanation for this lack of progress is simply that no theory has yet hit upon the correct set of invariant properties. Another explanation is that the goal of solving the lack of invariance problem is itself misguided. The primary aim of this paper is to suggest that the latter explanation is correct. The lack of progress in explaining speech perception exhibited by the current research framework, I'll suggest, is not, in the first instance, due to the failure of individual theories to solve the lack of invariance problem, but rather to the common background assumption that doing so *is* in fact the key to explaining speech perception.

My criticism of the underlying goal of speech perception research depends on noting that the current research framework has thus far been an empirical failure. Though the theories occupy a spectrum of methodological approaches and have amassed a great deal of indirect empirical support for their claims, the stark fact remains that most theories have not identified even a single candidate invariant property. The one theory for which this is not true -- the theory of acoustic invariance -- can produce such a candidate for only a single class of consonants. The extent and duration of this empirical failure make it reasonable to consider whether the failure is due primarily to the central question being posed rather than to the answers thus far offered.

To support this suggestion I argue that, in addition to generating extremely weak empirical results, the current research framework also generates theories which are weak. Current theories of speech perception are weak not only in the sense that virtually nothing in the world answers to their descriptions, but in the sense that the descriptions themselves are often ambiguous or even incoherent. Within the space of the present paper, I offer a brief analysis of the strategies of two theories of speech perception (the motor theory of speech perception and the theory of acoustic invariance) and try to show that their particular theoretical problems result from their common goal of trying to identify invariant properties of phonetic percepts.

2. THE STRATEGY OF THE MOTOR THEORY

The most basic claim of the motor theory (Liberman et al. 1967; Liberman and Mattingly 1985) is that the processes of speech

perception and production are closely linked. Speech is *special* according to the motor theory in that, of all the phenomena human beings perceive, speech sounds are the only ones that we also produce. Since we are not only perceivers of speech, but also producers thereof, we are said to have knowledge (in an appropriately tacit sense) of how speech sounds are produced.

The postulation of a link between perception and production figures in both major components of the motor theory. The motor theory claims (i) that phonetic percepts are identified with invariant units of production and (ii) that speech sounds are perceived in a specifically phonetic, as opposed to auditory, module -- a module mediated by production mechanisms. It's the first of these claims that I'll be primarily concerned with.

The objects of speech perception, according to the motor theory, are not to be found in the proximal stimulus -- the acoustic signal as it enters the ear -- but more nearly in the distal object -- the motor event that generates the acoustic signal. The objects of perception are meant to be the physical correlates of phonetic segments. They are described as "the physical reality underlying the traditional phonetic notions" (Liberman & Mattingly 1985: 2).

It is unclear, however, what these physical correlates are supposed to be; it is unclear, in other words, how the term "motor event" is to be construed in the motor theory of speech perception. The invariants of speech perception are to be identified with some level or other of the production process, but there's an equivocation in the motor theory over what particular level of the production process is supposed to contain the invariants. I'll suggest that this equivocation is not detachable from the central claims of the theory but is instead a central part of its strategy. That is, in addition to not being able to produce any candidate *empirical* invariants to support their theory, the motor theory of speech perception cannot even produce candidate *theoretical* invariants to describe their theory without relying on a systematic ambiguity.

The ambiguity is principally between the claim that neural commands are the invariants of phonetic perception, and the claim that articulatory movements are. Consider the following examples:

> The objects of speech perception are the...phonetic gestures of the speaker, represented in the brain as invariant motor commands that call for movements. (Liberman and Mattingly 1985: 2)

> [In the perception of phonetic structure,] the distal object is a phonetic gesture or, more explicitly, an 'upstream' neural command for the gesture from which the peripheral articulatory movements unfold. (Liberman and Mattingly 1985: 9)

> [T]he gestures do have characteristic invariant properties, as the motor theory requires, though these must be seen, not as peripheral movements, but as the more remote structures that control the movements. (Liberman and Mattingly 1985: 23)

Most immediately, the objects of speech perception are said to be neural objects: "invariant motor commands"; "an 'upstream' neural command'; "remote structures that control the movements". Yet they are not clearly distinguished from gross articulatory movements. In each of the quotes, articulatory movements are mentioned along with the neural invariants.

Still, whatever ambiguity exists regarding what the objects of speech perception are supposed to *be*, there is a distinct lack of ambiguity regarding what they are supposed to be *called*. The objects of speech perception in the motor theory are very clearly identified as *gestures*:

> Gestures are the objects of perception. (Liberman and Mattingly 1985: 10)

> First and fundamentally there is the claim that phonetic perception is perception of gesture. (Liberman and Mattingly 1985: 21)

> The invariants of speech perception are the phonetic gestures. (Liberman and Mattingly 1985: 29)

Now if the objects of perception are supposed to be neural structures, as the earlier quotes suggest most immediately, it is puzzling that they are regularly identified by a term that refers to gross articulatory movements. Although there is nothing to prevent a theory from taking an ordinary term and giving it a specialized meaning in a particular theoretical context, in the particular theoretical context of speech, 'gesture' *already* has a specialized meaning: it is the term reserved for movements of the articulators in the supralaryngeal vocal tract in the production of speech.

If, then, one were trying to distinguish neural commands from articulatory gestures of the vocal tract in order to specify the invariant objects of speech perception, one would not be aided in this effort by *labelling* the *neural* commands, *'gestures'*. Indeed, the confusion that appropriating this label would induce is so predictable that it is hard to see how it could be accidental.

But what do the motor theorists hope to accomplish by, as it were, marketing neural structures as articulatory structures? The explanation, I suggest, is as follows: The claim that articulatory structures are the invariant objects of phonetic perception is more *plausible* than the claim that neuromotor commands are. The objects of perception are what one perceives most directly in perceiving phonetic structure and it is radically implausible to claim that what one hears when perceiving speech is the sound of neurons firing.

By contrast, the claim that what one hears when perceiving speech is most immediately the articulatory gestures of the supralaryngeal vocal tract is not similarly implausible. For the articulatory gestures specify the dimensions and physical characteristics of the acoustic signal. The properties of the acoustic signal are the properties of the vibrating system that produced it. So it is more

plausible to claim that what one hears when perceiving speech is really the movements of the vocal tract, than to claim that what one hears, in an immediate sense, are neuromotor commands.

We can now begin to see how proponents of the motor theory hope to benefit from the articulatory associations of the term 'phonetic gesture' -- it is meant to lend plausibility or reasonability to their claim that the invariant units of phonetic perception are neuromotor commands.

But now the question arises, why the motor theorists don't simply drop the claim that neuromotor commands are the invariants of speech perception and claim explicitly that articulatory movements are. And here the reason seems clear: They don't do this because, although it is plausible to claim that there's an invariant mapping between phonetic percepts and articulatory structures, it is also *false*.

Speech is not produced by generating a series of static vocal tract shapes -- each one uniquely associated with a particular phonetic segment. Instead, individual vocal tract configurations are melded together in the production of speech; the articulators are in continuous transition from one target configuration to the next. Phonetic segments, in other words, are not articulated, but co-articulated. So, although the motor theorists would like to be able to claim that there is an invariant mapping between phonetic segment and articulatory shape -- in order to have a candidate solution to the lack of invariance problem -- because speech production is a dynamic process and because phonetic context influences the articulatory shape used to produce individual segments, they are unable to do so.

The ambiguity in the motor theorists' use of the phrase 'phonetic gestures', then, seems not to be a mere expository oversight. Rather it seems to have a substantive role to play in the theory. In order to insulate the theory from the charge of empirical inadequacy, phonetic gestures need to be understood as neural structures rather than articulatory ones; but, in order to insulate the theory from the charge of implausibility, phonetic gestures need to be understood in articulatory rather than neural terms. By leaving the referent of 'phonetic gestures' ambiguous between an articulatory interpretation and a neural one, proponents of the motor theory try to exploit the theoretical benefit of each interpretation, without incurring the theoretical burden of either.

3. THE STRATEGY OF THE ACOUSTIC INVARIANCE THEORY

Central to the theory of acoustic invariance (Blumstein and Stevens 1981; Stevens and Blumstein 1981) is the claim that invariant objects of speech perception are located directly in the speech signal. Proponents of this theory do not deny that the acoustic properties which have seemed to be invariant are, in fact, invariant. They do not claim, for example, that the formant frequency transitions widely believed to be context-dependent perceptual cues for stop consonants can, after all, be specified in context-independent terms. Instead, they deny that the display of the acoustic signal in which these variable properties are made salient -- frequency in the time domain as exhibited in spectrograms -- is the *only* display of the speech signal in which perceptually relevant properties are rendered salient. More precisely, they claim that acoustic properties of another sort exist which are both invariant with phonetic categories and used in the perception of speech.

The principal distinguishing characteristics of the acoustic properties which the current theory identifies as invariant are that they are integrated properties instead of individual ones, and that they are presented as frozen in time rather than as temporally extended. For example, acoustic information about the burst, the onset frequencies and the direction of the formant transitions of a stop consonant is presented as a single time-independent spectral display.

A complete theory of speech perception in this framework would need to specify invariant integrated properties sufficient to categorize all phonetic categories. While preliminary suggestions for a number of speech contrasts are discussed by proponents of the theory, the majority of the theoretical and empirical work in this framework has focussed on a single feature: place of articulation for stop consonants.

The invariant property for place of articulation in stop consonants posited by proponents of the acoustic invariance theory is *the gross shape of the spectrum sampled at (or near) the release of the closure*. The gross shape of the spectrum for bi-labial, alveolar, and velar stops can be characterized respectively as "diffuse falling", "diffuse rising" and "compact". The property of diffuseness refers to the "spacing" in frequency between successive peaks: diffuse peaks are relatively far apart as compared to compact spectra; in compact spectra two peaks may overlap. The property of rising or falling refers to the direction of successive peaks: whether there is a relative increase or decrease in amplitude with successive peaks. Compact peaks by contrast will tend to be dominated by one large central peak.

A number of objections might be raised regarding the strength of the evidence in favor of this theory (e.g., the quantitative evidence that the proposed invariant is in fact invariant, is not overwhelming; moreover, quantitative measures of proposed invariant properties are specified only for the single feature of place of articulation). However, I want to consider a more methodological objection here -- one which focuses less on the evidence the theory offers in its support, than on the evidence it discounts.

At the heart of this criticism is the claim that the basis for distinguishing the proposed invariant acoustic properties from other acoustic properties in the signal is problematic. The theory of acoustic invariance, nevertheless, *must* distinguish these two classes of properties because it claims that gross spectral shape sampled at the burst release is the primarily used acoustic property in the perception of speech. The difficulty, though, is that there

is a wealth of evidence that other acoustic cues are used to perceive these sounds.

The 2nd formant transition is perhaps the most obvious of these but many others have been identified as well (see Lisker 1978 for a list of 16 different cues thought to be used to distinguish [p] and [b] in medial position). Moreover, there is a growing consensus that 'every potential cue is an actual cue' -- that is, that *any* acoustic information for a given phonetic segment can, in the appropriate circumstances, be used to identify it. Acoustic cues are said to engage in "trading relations" (see Repp 1982): a loss of one cue can be compensated for perceptually by the enhancement of another. This suggests a kind of perceptual equivalence among acoustic cues, although in ordinary circumstances, some cues may well play a greater role than others.

The present criticism does not depend on denying this asymmetry of cues; it depends only on denying that this asymmetry always favors the acoustic properties identified by the theory of acoustic invariance. But -- and here is the problem -- we have no reason whatsoever to believe that this is so. The experimental data that the invariant acoustic properties identified by the theory are *used* in perception is weaker than the evidence that such invariants *exist*, but proponents of the theory offer no evidence at all that these properties are used *more*, or *more often* than the traditional (i.e., variable) acoustic properties.

Proponents of the theory of acoustic invariance do not deny the perceptual contribution of the variable acoustic cues; but they claim that their contribution to identifying phonetic percepts is perceptually subordinate. The point of the present criticism is that this claim is simply an empty stipulation; indeed, it is a stipulation which counters the existing empirical evidence.

What underwrites this central assumption of the theory of acoustic invariance seems to be, not the empirical evidence *per se*, but rather, an antecedent commitment to identifying phonetic percepts with invariant lower-level properties of the proximal stimulus -- that is, a commitment to a reductionist and context-independent approach to solving the lack of invariance problem. If one has such an antecedent commitment, one will then have a theoretical stake in claiming that the class of invariant acoustic cues plays a more important perceptual role than the class of variable acoustic cues. For in such circumstances it will be built in to the description of the goal of a speech perception theory that invariant acoustic cues are required to explain the process of speech perception. Now this strategy of basically privileging the class of evidence which confirms one's own theory is not *necessarily* pernicious. Whether or not it is, in a given case, will depend on how well motivated the theory's antecedent commitments and assumptions are. The difficulty in the present case is that it is the motivation for these assumptions and antecedent commitments themselves that is precisely being questioned. In such a case we can generate very little support for these assumptions by relying on claims whose only real justification is the truth of these assumptions themselves.

4. CONCLUSION

In the motor theory of speech perception, no empirical candidates for solving the lack of invariance problem are offered. The claim that such invariants exist is simply an empty stipulation of the theory. Even as a stipulation, though, the claim cannot be consistently maintained, for it depends on systematically equivocating between an articulatory and a neural interpretation of the invariant objects of speech perception. In the theory of acoustic invariance one empirical candidate for solving the lack of invariance problem is offered for a single feature. But accepting this candidate as the solution to the lack of invariance problem depends on the empty stipulation that a preponderance of contrary empirical evidence can be ignored. Although I have discussed only two examples of how the theoretical weaknesses of current speech perception theories can be tied to the underlying goal of trying to solve the lack of invariance problem, more examples can be generated. Taken cumulatively, along with the empirical failure of research in the current framework, it seems reasonable to conclude that progress in the overall goal of explaining how speech is perceived will be better served by redefining the central problem to be solved than by further efforts to resolve it.

REFERENCES

1. Blumstein, S.E. and Stevens, K.N. (1981). "Phonetic Features and Acoustic Invariance in Speech," *Cognition* 10: 25-32.

2. Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M. and Gerstman, L.J. (1952). "Some Experiments on the Perception of Synthetic Speech Sounds," *Journal of the Acoustical Society of America* 24: 597-606.

3. Liberman, A.M., and Mattingly, I. (1985). "The Motor Theory of Speech Perception Revised," *Cognition* 21: 1-36.

4. Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967). "Perception of the Speech Code," *Psychological Review* 74: 431-461.

5. Lisker, L. (1978). "Rapid vs. Rabid: A Catalogue of Acoustic Features That May Cue the Distinction," *Status Report on Speech Research*, SR-54: 127-32.

6. Repp, B.H. (1982). "Phonetic Trading Relations and Context Effects: New Experimental Evidence for a Speech Mode of Perception," *Psychological Bulletin* 92: 81-110.

7. Stevens, K.N. and Blumstein, S.E. (1981). "The Search for Invariant Acoustic Correlates of Phonetic Features," in P.D. Eimas and J.L. Miller, eds., *Perspectives on the Study of Speech*. New Jersey: Erlbaum.