Traunmüller, H. (1987) Phase vowels. In: M.E.H. Schouten (ed.) *The Psychophysics of Speech Perception*, Dordrecht: Martinus Nijhoff Publishers, 377-384.

Willems, L.F. (1966) The intonator. *IPO Annual Progress Report, 1*, 123-125.

aus Elsendoorn & Bouma 1988
Working models of human perception

Phonetic Invariance and the Adaptive Nature of Speech

Björn E.F. Lindblom*

'...after all planes do not flap their wings'

1 Introduction

My topic is a classical problem in phonetics and speech research: That of reconciling the physical and the linguistic descriptions of speech. Investigating speech we continually battle with the variability of the speech wave and hope for information and insights that will tell us how acoustic properties are related to 'features', 'segments' and other categories that we use in our linguistic analyses.

On the one hand, it is clearly true that in spite of several decades of acoustic phonetic research on many languages, we still encounter serious difficulties when it comes to specifying phonological units in such a way that their phonetic description will remain invariant across the large range of contexts that the communicatively successful real-life speech acts present to us.

On the other hand, it is also true that many of us share the conviction that taking steps towards the solution of the invariance problem will be crucial to acquiring a deeper theoretical understanding of human speech as well as to developing more advanced systems for speech-based man-machine communication (Perkell and Klatt, 1986).

I have organized my presentation in terms of three questions: Is phonetic invariance articulatory? Is it acoustic? Or is it auditory? Let us begin by reviewing some experimental findings that appear to identify phonetic aspects that remain constant although the speech analysed undergoes various transformations.

* Department of Linguistics, Stockholm University, S-106 91 Stockholm, Sweden, and University of Texas at Austin, Austin 78712-1196, Texas, USA.

2 Is phonetic invariance articulatory, acoustic or auditory?

In their recently revised statement of the motor theory of speech perception Liberman and Mattingly (1985) draw attention to one of the assumptions that underlie their theory: *"Phonetic perception is perception of gesture..."* and *"the invariant source of the phonetic percept is somewhere in the processes by which the sounds of speech are produced"* (p.21). The authors dwell on the variability that articulatory gestures tend to exhibit in instrumental phonetic records but they maintain that *"it is nonetheless clear that, despite such variation, the gestures have a virtue that the acoustic cues lack: instances of a particular gesture always have certain topological properties not shared by any other gesture"*. In conclusion they argue that *"the 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. These structures correspond to the speaker's intentions"*.

Comparing phonetic invariants to the speaker's intentions Liberman and Mattingly remind us of Baudouin de Courtenay's (1845-1929) pioneering definition of a phoneme as 'eine Lautabsicht' and of the more recent conceptualizations that fall under the heading of the so-called target theories of speech production. In the sixties such theories were explored by several investigators in the hope that a lot of the variability that speech signals typically exhibit – e.g., vowel-consonant coarticulation (Öhman, 1967) – could be explained in terms of the spatial and temporal overlap of adjacent 'motor commands' (MacNeilage, 1970). Articulatory movements were seen as sluggish responses to an underlying forcing function which was assumed to change, usually in a step-wise fashion, at the initation of every new phoneme (Henke, 1966). Owing to variations in say stress or speaking tempo different contexts would give rise to differences in timing for a given sequence of phoneme commands. Articulatory and acoustic goals would not always be reached, a phenomenon termed 'undershoot' by Stevens and House (1963). But since such undershoot appeared to be lawfully related to the duration and context of the gestures, the underlying articulatory 'targets' of any given phoneme – 'die Lautabsicht' – would nevertheless, it was maintained, remain invariant (Lindblom, 1963).

Duration-dependent undershoot, as proposed by ourselves in the sixties, is schematized in figure 1. The model predicts the formant frequencies of a vowel as function of its identity, its consonantal environment and its

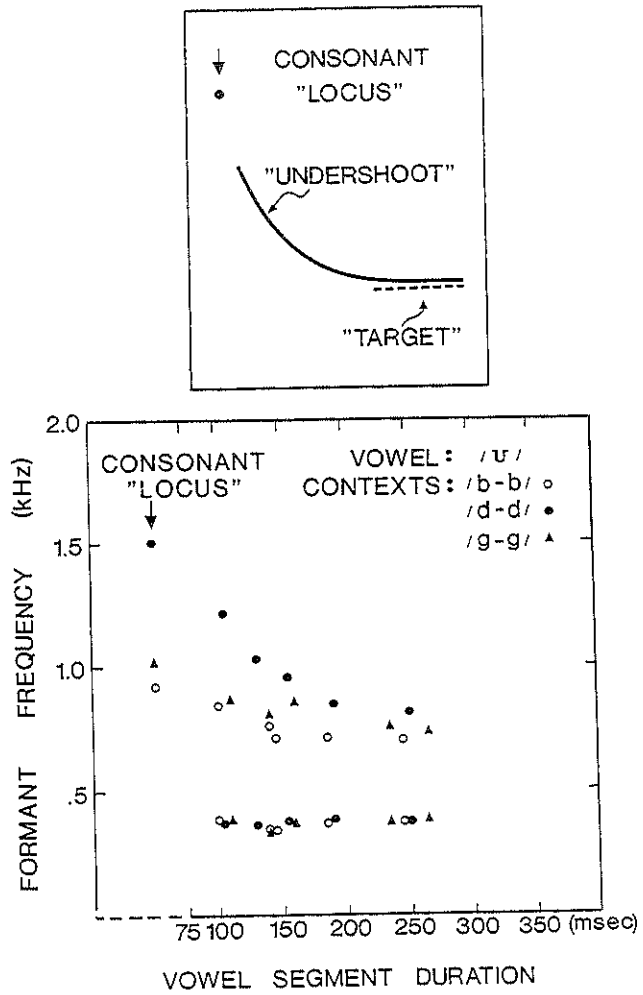

Figure 1: Data from Lindblom (1963) and the 'undershoot' model, an exponential duration- and 'locus'-dependence of vowel formant frequency displacement in CVC syllables.

duration. For biomechanical reasons undershoot still seems to be a phonetically valid notion. It captures a real constraint on speech production. However, in the light of more recent evidence it has become clear that this model represents the notion of target much too simplistically and secondly, undershoot is not as inevitable a phenomenon as the model implies.

To illustrate some of the problems, let us first examine how the idea of an invariant target fares in a simple experiment in which subjects are instructed to vary their degree of vocal effort.

The measurements in figure 2 are based on three Swedish subjects say-

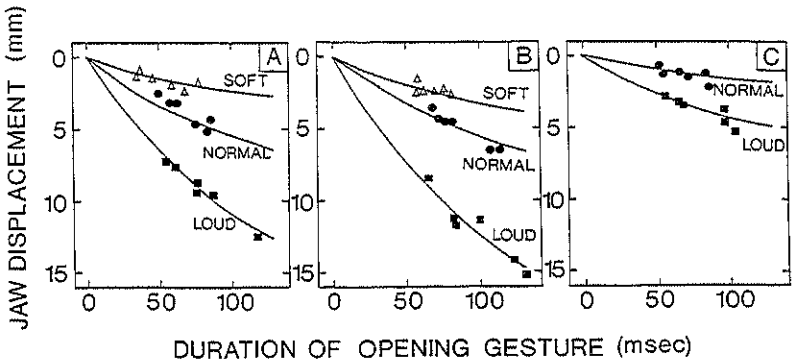

Figure 2: Extent of peak jaw displacement as function of the duration of the jaw opening gesture. Three speakers: A, B, C. Three degrees of vocal effort.

ing the nonsense words *'dadd, daddad, dadadd, daddadad, dadaddad'* and *'dadadadd'* at three degrees of loudness: softly, normally and loud. All observations pertain to the stressed vowel which, orthographically, is the one followed by the double consonant. The broad range of durational values along the x-axis were obtained by systematic variations in word length and position within the words (Lindblom, Lubker, Lyberg, Branderud, and Holmgren, 1987). Mean peak jaw displacement is plotted against gesture duration for the stressed vowel. In each panel the curves and the associated data points represent the three degrees of vocal effort. Although the expected exponential duration-dependence appears in all cases it seems impossible to claim that there is a single speaker-specific articulatory target, a unique, invariant jaw position underlying the three effort conditions.

Information indicates that in fast speech articulatory and acoustic goals can be attained despite short segment durations (Engstrand, *to appear*;

Gay, 1978; Kühn and Moll, 1976). Moreover, undershoot has been demonstrated in unstressed Swedish vowels that exhibit long durations owing to 'final lengthening' (Nord, 1986). Such exceptions from simple duration-dependence appear to highlight the reorganizational abilities of the speech production system. One way of resolving the problem posed by these somewhat contradictory results might be obtained if it were shown that when instructed to speak fast subjects have a tendency to 'overarticulate', thus avoiding undershoot to some extent, whereas when destressing they are more likely to 'underarticulate' (cf. discussions below of hypo- and hyper-speech). Also note the possibility of language-specific patterns of vowel reduction which becomes particularly relevant when addressing such questions (cf. Delattre's (1969) discussion of English, French, German and Spanish).

In summary, the original observations of 'undershoot' carried the implication that the invariant correlates of linguistic units were to be found, not in the speech wave nor at an auditory level, but upstream from the level of articulatory movement. Phonetic invariance was accordingly associated with the constancy of underlying 'spatial articulatory targets' (for reviews of the target concept see e.g., MacNeilage, 1970, 1980). However, subsequent experimentation – some of which we already hinted at above – has revealed that the notion of segmental target must be given a much more complex interpretation.

Studies of compensatory articulation reinforce this conclusion particularly strongly. Let us summarize some results from an experiment using the so-called 'bite-block' paradigm (Lindblom, Lubker, Lyberg, Branderud, and Holmgren, 1987). We asked native Swedish speakers to pronounce mono-, bi- and trisyllabic words under two conditions: normally and with a large bite-block between their teeth. They were instructed to try to produce the bite-block utterances with the same rhythm and stress pattern as the corresponding normal items. Real Swedish words as well as 'reiterant' nonsense forms were used: *'Bob – babb, Bagdad – babbab, va snabb! – bababb, Waldemar – babbabab, begabbad – bababbab, falla dagg – bababbb'*. Measurements were made of the acoustic durations of the consonant and vowel segments of the reiterant speech samples. By comparing the normal and the bite-block versions we wanted to address the question whether subjects would be able to achieve the bilabial closure for the /b/ segments in spite of the abnormally low and fixed jaw position and whether they would be able to do so reproducing the normal durational patterns.

The diagram to the right in figure 3 compares stressed vowel durations for normal and bite-block conditions. It is representative of our finding that

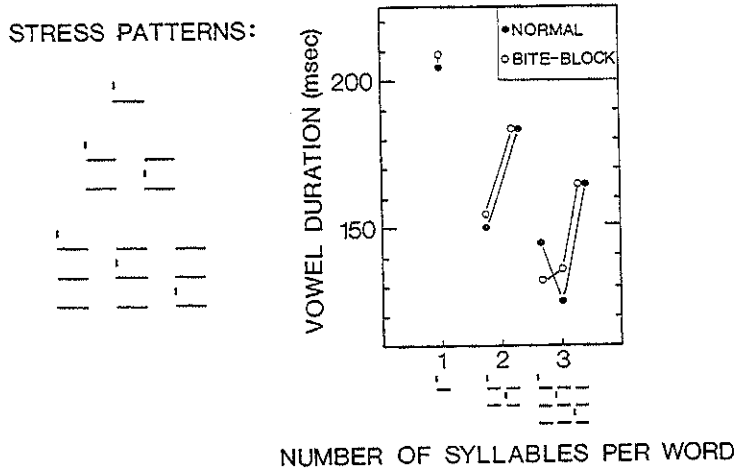

Figure 3: Comparison of normal speech and compensatory 'bite-block' articulations. Average stressed vowel segment duration in nonsense words with stress patterns as indicated in the left part of the diagram.

the timing in the bite-block words deviated systematically but very little from the normal patterns. These results enable us to conclude that our subjects were indeed capable of compensating. To explain the results it appears reasonable to suggest that a representation of the 'desired end-product' – the metric pattern of the word – must be available in some form to the subjects' speech motor systems and that successful compensation implies a reorganization of articulatory gestures that must have been controlled by such an output-oriented target representation. These results are in agreement with those reported earlier by Netsell, Kent and Abbs (1978). Moreover, they are completely analogous to the previous demonstrations that naive speakers are capable of producing isolated vowels whose formant patterns are normal at first glottal pulse in spite of an unnatural jaw opening imposed by the use of a 'bite-block' (Lindblom, Lubker and Gay, 1979; Gay, Lindblom and Lubker, 1981).

These results bear on the recent discussion of speech timing as 'intrinsically' or 'extrinsically' controlled. Proponents of action theory (Fowler, Rubin, Remez and Turvey, 1980) approach the physics of the speech motor system from a dynamical perspective with a view to reanalysing many of the traditional notions that now require explicit representation in extant speech production models such as 'feedback loop', 'target', etc. Their writ-

ings convey the expectation that many aspects of the traditional 'translation models' will simply fall out as consequences of the dynamic properties intrinsic to the speech motor system. In the terminology of Kelso, Saltzman and Tuller (1986, p.55) *"... both time and timing are deemed to be intrinsic consequences of the system's dynamical organization"*. Methodologically, action theory is commendable since, being committed to interpreting phonetic phenomena as fortuitous (intrinsic) consequences rather than as controlled (extrinsic) aspects of a speaker's articulatory behaviour, it guarantees a maximally thorough examination of speech production processes. However, it is difficult to see how, applying the action theoretic framework to the data on compensatory timing just reviewed, we would possibly avoid postulating some sort of 'temporal target' representation which is (1) extrinsic to the particular structures executing the gestures and which is (2) responsible for extrinsically tuning their dynamics.

The point to be made here is that speech production is a highly versatile process and sometimes appears strongly listener-oriented.

The plasticity of the speech motor system is further illustrated by what might be called a 'natural bite-block' experiment recently done by Richard Schulman in our Stockholm Laboratory (Schulman, forthcoming). Schulman compares syllables in loud and normal speech and observes, as others have done, that loud vowels have a more open jaw position than those of normal syllables. Figure 4 demonstrates the fact that Schulman's subjects were found to use loud jaw openings in vowels that were more or less uniformly three times larger than normal ones.

Two observations call for special comments. The first concerns the formant patterns of loud vowels. We shall return to that in a moment. The second is related to an associated result shown in figure 5 which compares loud vowel durations along the ordinate with the corresponding normal values on the abscissa: We see that, relatively speaking, vowel durations in loud speech are longer whereas loud consonant durations are shorter. This is in agreement with previous results reported by Fónagy and Fónagy (1966).

What does that result mean? The normal-loud vowel duration differences look suspiciously similar to the durational differences between normal open and close vowels which have been observed for many languages (Lehiste, 1970). Finding that the duration of the EMG recorded from the anterior belly of the digastric correlated with both mandibular displacement and vowel duration, Westbury and Keating (1980) have suggested that this temporal variation among vowels, although non-distinctive, must be seen as present in the neuromuscular signals controlling their articulation. An

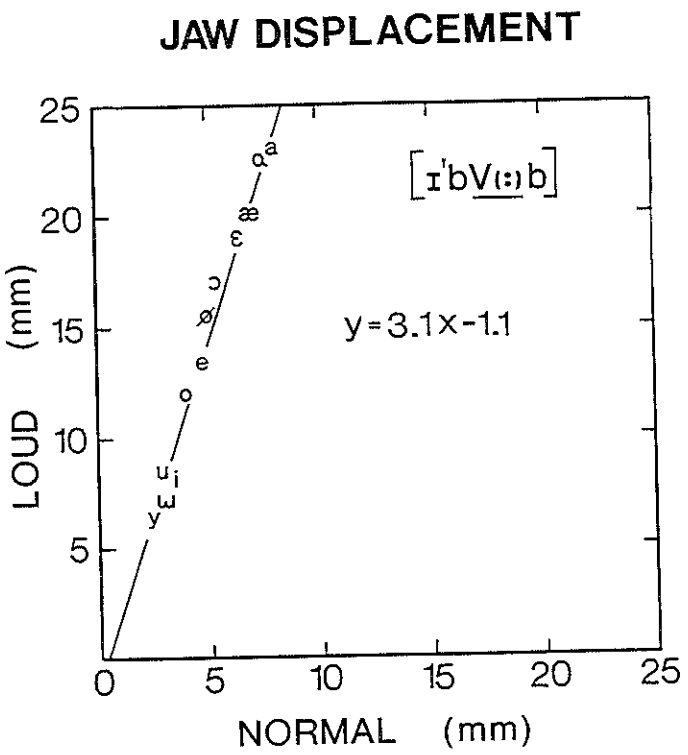

Figure 4: Comparison of jaw positions for Swedish vowels in loud and normal syllables. Average data for Swedish vowels from subject HH (adapted from Schulman, forthcoming).

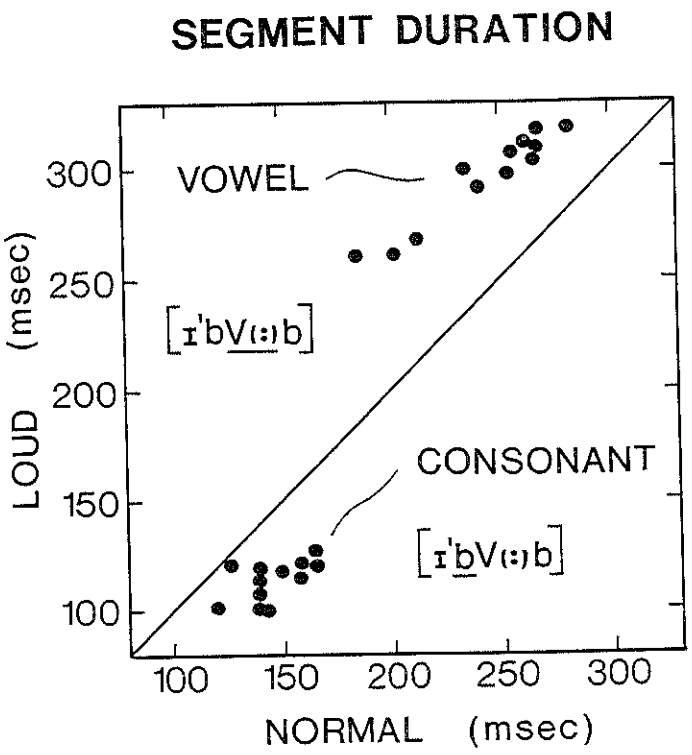

Figure 5: Comparison of vowel and consonant durations in loud and normal speech. Average data for Swedish subject (adapted from Schulman, forthcoming).

alternative interpretation would be to invoke Eli Fischer-Jørgensen's 'extent of movement hypothesis' (1964) which attributes the longer duration of open vowels to the circumstance that a more extensive jaw lowering causes a premature release of the opening gesture and delays the closing gesture.

The question whether the open-close vowel duration difference is an intrinsic or extrinsic phonetic phenomenon is accordingly somewhat controversial. Schulman's findings bear on the problem. He constructed a model of loud speech based on the observation that loud movements appear to be 'exaggerated' versions of the corresponding normal gestures. That effect is illustrated by figure 6 which compares the vertical separation of the lips

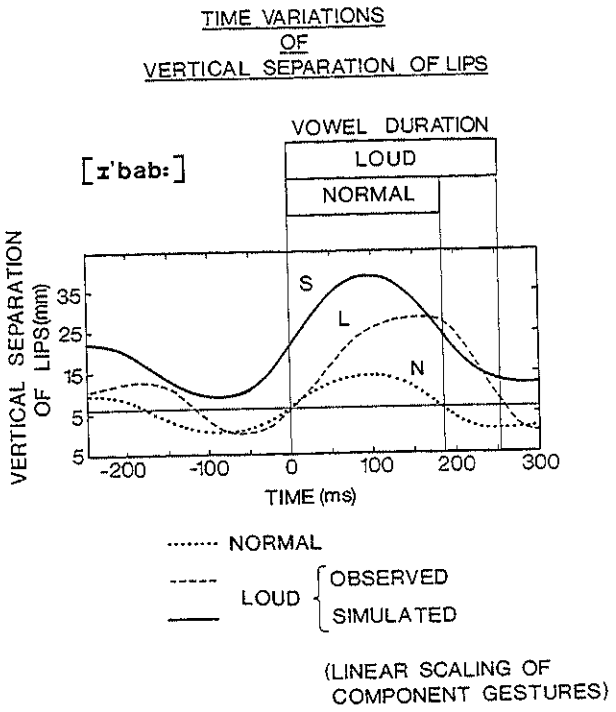

Figure 6: Time variations of vertical separation of the lips in loud (L) and normal (N) productions of test word. Also shown is a curve labelled S derived by linear scaling of the normal component gestures of vertical lip separation, that is vertical displacements of the upper and lower lips and jaw (adapted from Schulman, forthcoming).

in the test word /iba:b/ for the two experimental conditions (the N- and L-labelled curves). The horizontal line indicates the value of this parameter

at which the mouth opens and closes. Accordingly it provides the criterion for determining where the vowel segment begins and ends respectively.

Schulman attempted a simulation of the loud condition by postulating that the lips and the jaw are linear mechanical systems and that loud differs from normal speech solely in terms of the amplitudes of the underlying excitation forces. In accordance with that assumption a linear scaling was performed of the normal articulatory movements that determine the vertical separation of the lips. In other words, simulated vertical lip separation was derived by a summation of the linearly scaled versions of vertical displacements of the upper and lower lips and the jaw (S-curve). By using the criterion mentioned in the previous paragraph for determining vowel segment duration he was then able to predict the durations of vowel and consonant segments for loud speech. He found that linear scaling produced much too long vowels, or as indicated here by the S-curve, eliminated stop closures entirely. We conclude from his heuristic exercise that the 'loud transform' cannot be described as a simple scaling of component gestures but entails extensive goal-oriented reorganization of articulatory movements.

The implication of this result is that it attributes the durational differences to a superposition effect, thus supporting Fischer-Jorgensen's 'extent of movement hypothesis'. Schulman concludes that, unless the effect of opening and closing of the jaw had been actively counteracted, loud and normal vowel durations would have differed even more than they actually did.

In summary, the preliminary implication of all work touching the theme of compensatory articulation appears to be that – whether we use 'target' with reference to segmental attributes, segment durations or patterns of speech rhythm – the term is better defined, not in terms of any simple articulatory invariants, but with respect to the acoustic output that the talker wants to achieve. If phonetic invariance is not articulatory could it be acoustic then? The results reviewed so far thus point in the direction of our second question.

The suggestion that the speech signal contains absolute physical invariants corresponding to phonetic segments and features has received a lot of attention thanks to the work by Stevens and Blumstein (Stevens and Blumstein, 1978, 1981; Blumstein and Stevens, 1979, 1981). The idea has been favourably received by many, for instance by Fowler in her attempts to apply the perspective of 'direct perception' to speech (Fowler, 1986).

Others have been provoked to emphasize the inadequacy of the non-dynamic nature of the template notion (Kewley-Port, 1983) and the sub-

stantial context-dependence that the stop consonants of various languages typically display even in samples of carefully enunciated speech (Öhman, 1966).

Incidentally, let us note that, if it exists, acoustic invariance is a rather strange notion since talkers can only monitor it through their sences and listeners can only access it through their hearing system. Why should sensory feedback and auditory transduction be assumed to impose negligible transformation of the acoustic signal? Is it the case that what people really mean when they talk about acoustic invariance if in fact 'auditory' invariance?

To introduce our third question: 'Is phonetic invariance auditory?' Let us comment on the second finding in Schulman's loud speech study. Since loud vowels show greater jaw openings, and since lowering the jaw is known to raise markedly the first formant frequency (Lindblom and Sundberg, 1971), the question arises: Do subjects compensate for the greater jaw opening the way they do in the bite-block experiments (Lindblom *et al.*, 1979, Gay *et al.*, 1981)? In other words, do formant patterns remain invariant?

The answer is no. The first formant of loud vowels is shifted upwards by about one Bark. This result offers a rather curious parallel to an observation made by Hartmut Traunmüller in our laboratory (Traunmüller, 1985). The so-called 'Traunmüller effect' is reminiscent of Sundberg's findings on $F_0 - F_1$ interrelationships in soprano vowels (Sundberg, 1975). It is a demonstration of the transforms required to preserve the perceptual constancy of vowel quality under changes in (1) vocal effort and (2) vocal tract size.

Effort and vocal tract variations can be dramatically illustrated by synthetically modifying a naturally spoken /i/. When all formants and F_0 are shifted equally along a Bark scale an /i/-like vowel is perceived but the voice changes from an adult's to a child's. When both F_1 and F_0 are varied in such a way that $F_1 - F_0$ is kept constant on a Bark scale – and the upper formant complex is left unchanged – an /i/-like vowel is perceived. This is remarkable in view of the fact that F_1 reaches a value more typical of a low-pitched /ae/. One's impression is that the speaker remains the same but that she 'shouts'.

Note specifically the acoustic conditions for maintaining consonant quality under variations in vocal effort. They consist in shifting F_1 and F_0 *en bloc* and by dissociating these parameters from the upper formant complex. From the viewpoint of our current models of vowel perception this effect – the 'Traunmüller effect' – must be said to be a rather novel and unexpected phenomenon.

Also note the parallel between Schulman's and Traunmüller's results.

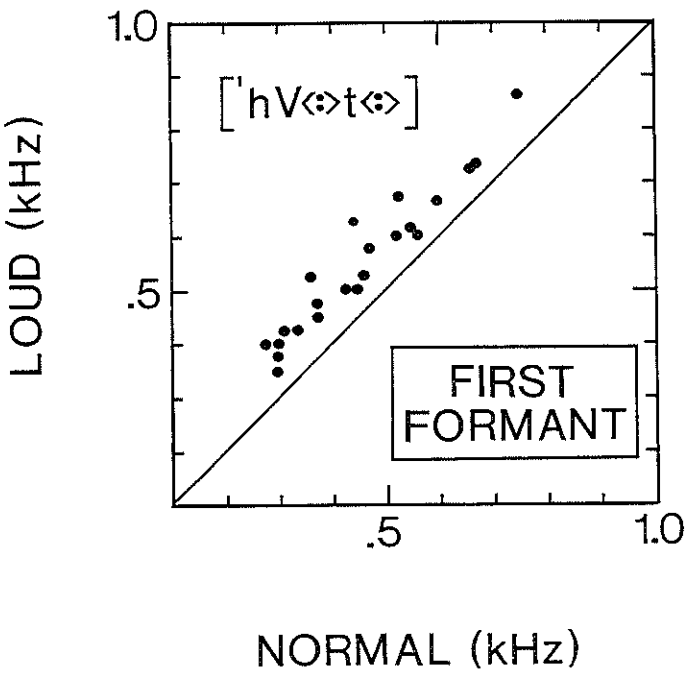

Figure 7: Comparison of first formant frequency for loud and normal speech. Average data for Swedish vowels from subject BG (data supplied by Schulman).

Are the findings causally related? Do we explain the lack of formant compensation in loud speech in terms of the 'Traunmüller effect'? Or do we account for the vowel quality results in terms of the 'Schulman effect'?

Of importance for the present discussion is the fact that behavioural constancies have been demonstrated and that they imply that at least in this case phonetic invariance is present at a level of auditory representation.

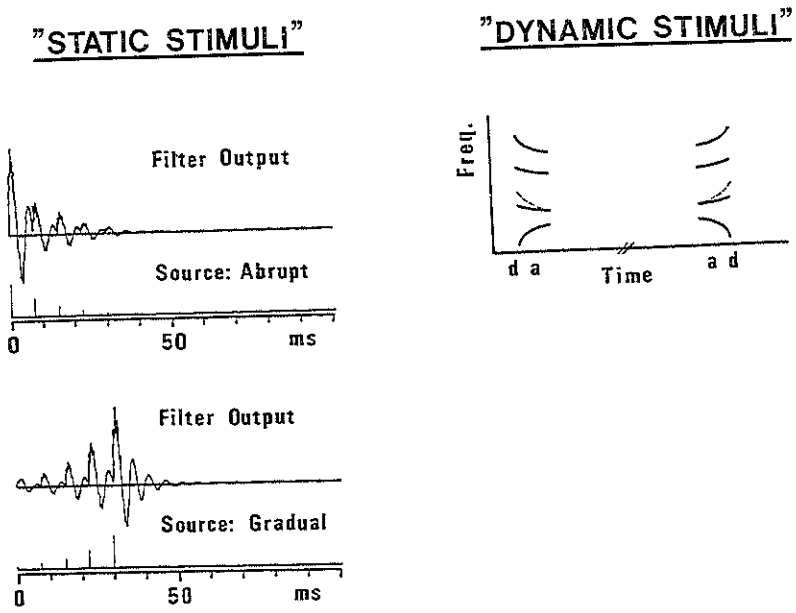

Figure 8: Stimuli used in F_2 difference limen study: short tone burts with static formant patterns and speech-like /da/- and /ad/-syllables (Lacerda, 1986).

Another set of observations suggesting some form of auditory invariance comes from work by Francisco Lacerda also from Stockholm University (Lacerda, 1986, 1987*a*, 1987*b*). We can characterize one part of his research as variations on the theme struck by Flanagan in his early 'difference limen' experiments on vowel formant frequencies (Flanagan, 1955).

Lacerda's question was: How well can listeners discriminate four-formant stimuli that differ solely in terms of the frequency of F_2. His work permits us to compare a psycho-acoustic task with a 'speech task': The discrimination of F_2 in brief tone bursts in which formant patterns are static (left part of figure 8) and the discrimination of the onsets and offsets of F_2 in speech-like stimuli, that is in the dynamic F_2-transitions in /da/- and /ad/-stimuli

(right part of figure 8).

Some of the results are shown in figure 9. They indicate that the sub-

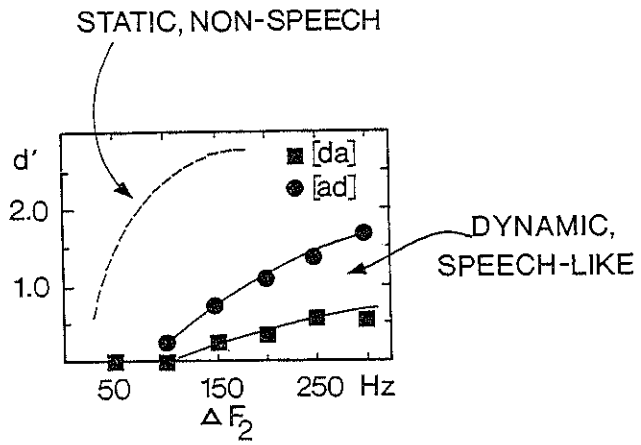

Figure 9: Results of F_2 difference limen study: F_2 differences are discriminated better in short tone burts with static formant patterns than in speech-like /da/- and /ad/-syllables with formant transitions (Lacerda, 1986). The tone burst results are in close agreement with Flanagan's 1955 data.

jects perform much better on the psycho-acoustic task (dashed curve) than on the speech-like test (solid lines). One might want to argue that Lacerda has here encountered the old fact first reported by Liberman and his colleagues at Haskins Laboratories as early as in the fifties, viz. the fact that intra-phonemic discrimination is considerably worse than inter-phonemic discrimination (Liberman, Harris, Hoffman and Griffith, 1957). Since the speech-like stimuli all fall within the /d/-category the results merely reflect the absence of a 'speech mode' mechanism for the psycho-acoustic stimuli and the presence of one for /da/ and /ad/. Others might prefer the possibility that the listeners' responses are due to an interaction between dynamic stimulus properties and speech-independent auditory processing (Lacerda, 1987*b*). I shall not discuss these and other interpretations further.

Rather my main point is this: The relatively large difference limens observed for speech-like stimuli indicates that percepts tend to be constant in spite of rather drastic variations in formant transition onsets. Such constancy is not acoustic: it clearly presupposes auditory processing.

3 Systematic nature of phonetic variability: the hyper-hypo dimension

Anyone who has made carefully auditory analyses and transcriptions of spontaneous speech will testify that its physical contents – as revealed say by spectrograms – is often full of surprises. Omissions and contextual modifications – both phonetic and phonological – tend to escape even the trained ear. These effects are sometimes referred to as elliptical speech.

Figure 10 is from our current investigation of different speaking styles in

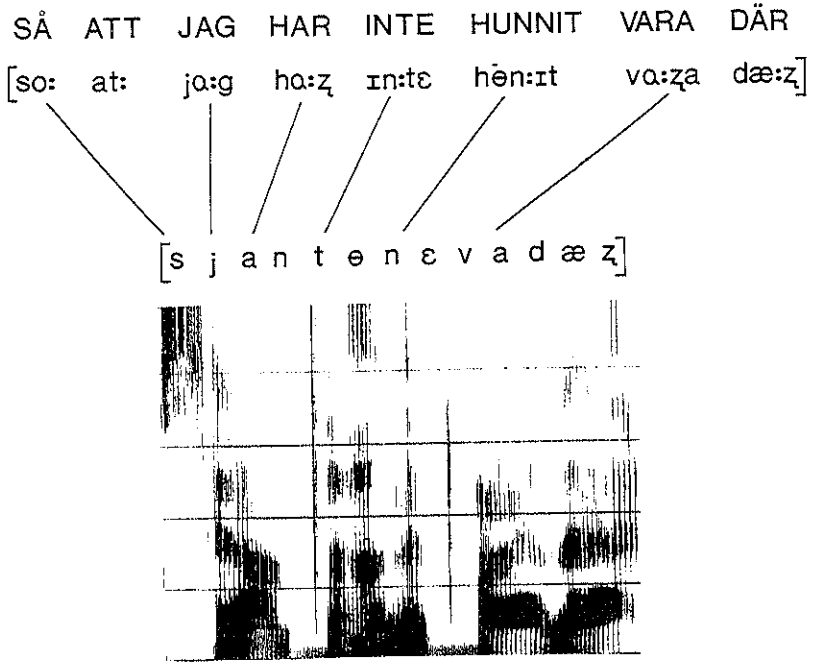

Figure 10: Phonetic transforms in different speaking styles. Spectrogram and narrow transcription of an example of 'elliptical speech', a Swedish utterance excerpted from spontaneous conversation: 'Så jag har inte hunnit vara där'. Also shown transcriptions of the words of the utterance in citation form pronunciation.

Swedish (Lindblom and Lindgren, 1985). So far our method has simply been to compare spontaneous and so-called 'self-generated speech' with readings of the same samples from lists. The first line of figure 10 is an orthographic representation of the Swedish utterance: 'Så jag har inte hunnit vara där'.

The second line gives the citation form pronunciations of the individual words. The third line is a more narrow transcription of a sample on the spectrogram which is an utterance recorded in a spontaneous conversation. Note the drastic changes between the styles. In our experience reductions and transforms of this type are common in casual Swedish. It would be surprising if such effects were limited to Swedish.

What are the implications of variations in speaking style for the invariance issue? Everyday experience indicates that speaking is a highly flexible process. We are capable of varying our style of speech from fast to slow, soft to loud, casual to clear, intimate to public. We speak in different ways when talking to foreigners, babies, computers and hard of hearing persons. And we change our pronunciation as a function of the social rules that govern speaker-listener interactions (Labov, 1972).

We recently undertook a literature survey[1] in order to systematize the types of speech materials that have been used in acoustic phonetic studies published during the past ten years. A total of over 700 articles were selected as preliminarily relevant. We ended up choosing 216 as meeting our criterion of 'descriptive study of speech based on quantitative acoustic phonetic measurements'.

Of special interest to us was to ascertain the relative proportions of studies investigating 'self-generated' speech such as spontaneous conversation on the one hand, and speech samples chosen by the experimenter such as list readings, nonsense words and so forth on the other. Not surprisingly, we found that the majority of studies, over 90%, use speech samples directly controlled by the experimenter.

One way of justifying this widely used procedure is to argue that first we will solve the problem of phonetic invariance in 'lab speech'. Then we will get to work on 'natural speech'. Another line of reasoning is to suggest that, although we lack the supplementary methodology required by 'ecological' speech, the excessive use of 'lab speech' introduces an undesirable bias in our data bases as well as in our theoretical intuitions about invariance and other key issues. And this bias might make us underestimate the problem of speech variability in spite of the fact that it is readily acknowledged by all workers in the field and has already, it would appear, been rather massively documented. Consequently the situation ought to be balanced.

I would like to present a few more illustrations of the effects of style on

[1] I am indebted to Diana Krull for doing the preliminary selections and to Natasha Beery of the Phonology Laboratory, University of California, Berkeley for the statistical analyses.

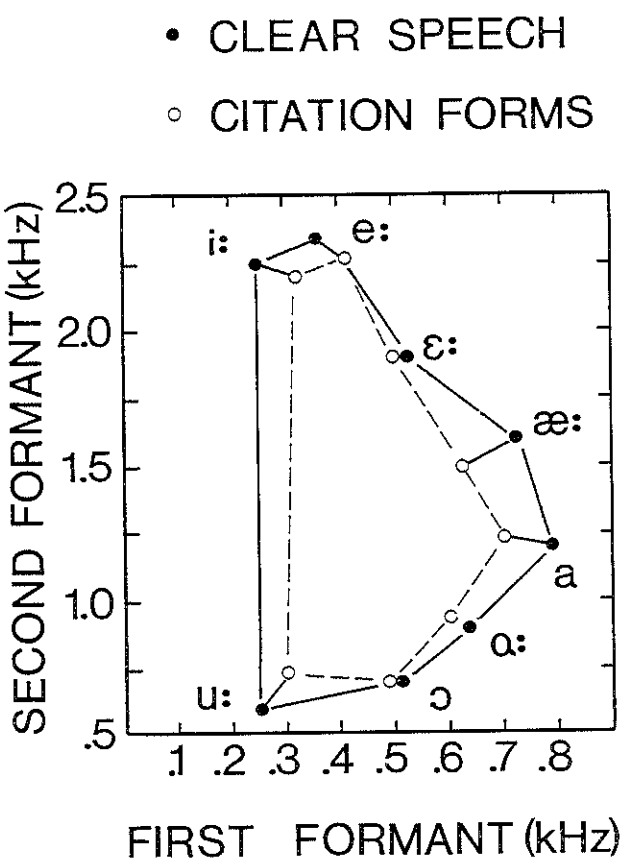

Figure 11: Swedish vowel formant frequencies in syllables spoken in response to request for clarification and in neutral list reading.

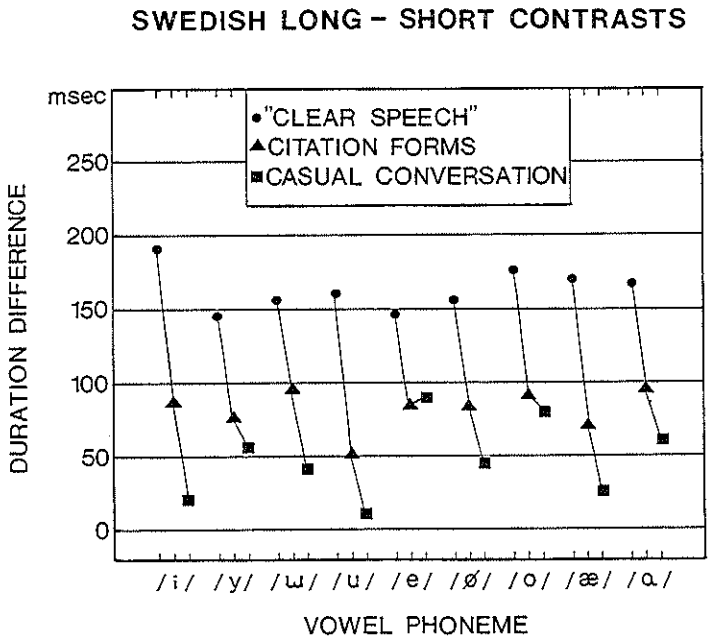

Figure 12: Durations of Swedish vowels in three speaking styles.

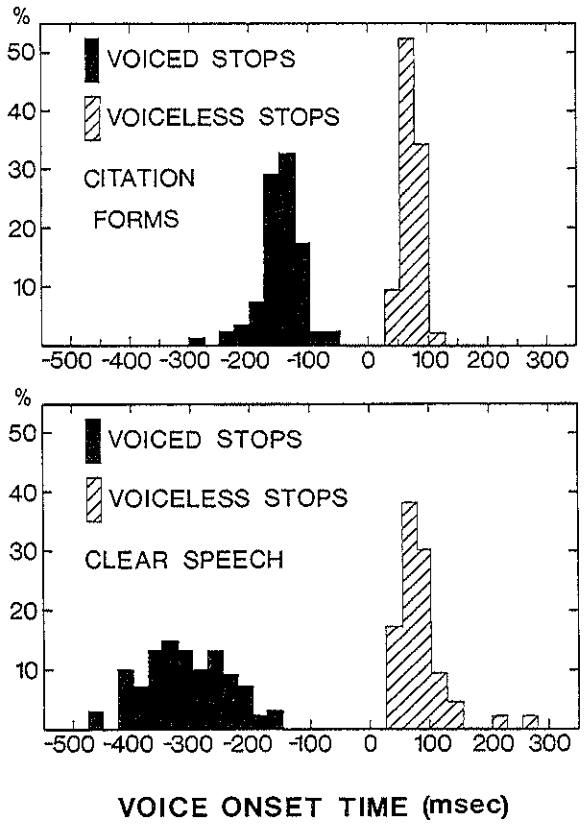

Figure 13: Voice onset times for Swedish voiced and voiceless stops in syllables spoken in response to request for clarification and in neutral list reading.

the phonetics of Swedish. For instance, we have found that the vowel space shrinks in casual style and is expanded in 'clear, hyperspeech' modes in relation to citation form pronunciations (figure 11). Durational differences between long and short vowels tend to decrease as we go from clear to neutral and casual styles (figure 12). Contrast in voice onset time for voiced and voiceless stops increases as we compare clear speech with neutral reference pronunciations (figure 13).

Figure 14 demonstrates a method of quantifying the extent of consonant-

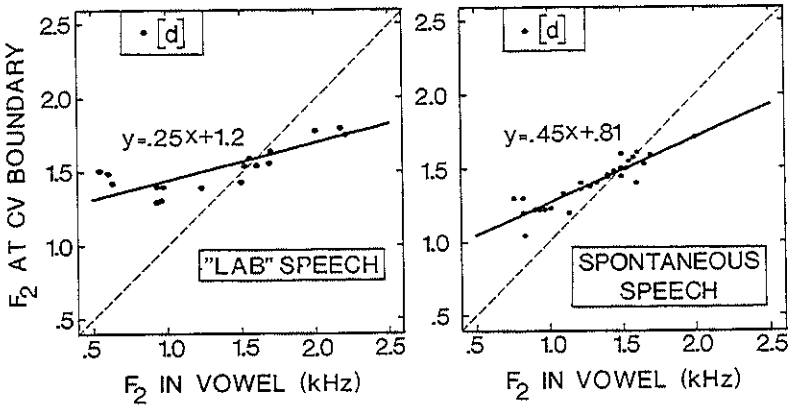

Figure 14: 'Locus'-equations for CV-sequences: citation form readings of test-words ('lab speech') and samples excerpted from running informal conversation ('spontaneous speech'). Ordinates: Value of F_2 at CV-boundary. Abscissa: Value of F_2 at vowel midpoint of corresponding syllable.

vowel coarticulation in the form of approximately linear equations – or, with a slight modification of Haskins terminology, 'locus'-equations. These diagrams are constructed by plotting formant frequencies at consonant-vowel boundaries as a function of adjacent vowel formant frequencies. Acoustic theory indicates that for most consonant-vowel combinations near-linear relationships should be expected.

As shown here, locus equations show a smaller slope for citation forms than for spontaneous speech. Since it is the case that the more horizontal the line the less the so-called 'locus' depends on the vowel, we are justified in concluding that vowel-consonant coarticulation seems to be counteracted in listener-orientated, clear styles but tends to be tolerated in casual speech.

Although I am unable to make my present point on the basis of a large body of quantitative acoustic data from several languages, the observations

we have made so far suggest that the prospects for any strong version of physical invariance to be substantiated across a large range of speaking styles seem most unfavourable.

4 How do we account for these experimental findings?

How do we account for all the diverse facts just reviewed? And how do we use them to formulate our reply to the three initial questions? Is phonetic invariance articulatory, acoustic or auditory?

Let us consider two major possibilities: The first is implicit in the manner in which we frame the questions. It presupposes that invariance is to be defined at the level of the signal. It says, in other words, that the invariance issue is indeed a problem of signal analysis. A strong version of this view would be to define the invariant correlates of linguistic units in terms of absolute acoustic properties invariably present in the speech wave. Another version might direct our search for invariance towards constant relations of higher order among acoustic properties. For instance, although highly variable in the absolute sense, the acoustic attributes of a given unit would still exhibit invariance if their status relative to other signal information remained constant.

The second possibility states in fact that we have asked our three questions the wrong way. Invariance cannot be seen as a phonetic problem. It is not a signal analysis problem at all. For the invariance of linguistic categories is ultimately to be defined only at the level of listener comprehension.

We have been persuaded by this latter point of view. It appears to give us the means for presenting a unified account of a broad spectrum of experimental phonetic findings including the ones just reviewed.

Let us expand on this claim in the context of a few brief comments on perceptual processes. At Stockholm University a test is currently used to measure how proficient native Swedish students are in understanding spoken French (Stöök, unpublished results). The task of the students is to listen to triads of stimuli consisting of two identical sentences and one minimally different and to indicate the odd case. Examples of test sentences are:

Montre-leur ce chapeau, s'il te plaît
Montre-leur ce chapeau, s'il te plaît
Montre-leur ces chapeaux, s'il te plaît

and

Je veux manger une soupe à l'oignon
Je veux manger une soupe à l'oignon
Je vais manger une soupe à l'oignon

Native speakers of French have no problems with such sentences whereas Swedish listeners knowing no or little French have a lot of trouble. However, when the key information – e.g., the ce/ce/ces or veux/veux/vais triads – is presented as fragments gated from the original sentences the performance of the Swedish subjects improves radically (Dufberg and Stöök, unpublished results).

This test can serve to remind us that perception is a product of two things: signal-dependent and signal-independent information, the latter including language-driven processing. While I am perfectly capable of discriminating the French minimal contrasts as auditory patterns I would quickly lose those patterns in a sentence context unless I have a sufficiently good command of French – that is access to signal-independent 'knowledge' whose interaction with the signal is a part of forming the final percept.

The speech literature is full of experimental data indicating that processes not primarily driven by the signal play an important role in the perception and understanding of speech. Such a conclusion can be supported by results from numerous paradigms including intelligible speech despite noise and distortion, improvement of identification as the signal gets linguistically and phonetically richer (Pollack and Pickett, 1964), detection of mispronunciations (Cole, 1973), word frequency effects (Luce, 1986), restoration (Warren, 1970; Ohala and Feder, 1986), fluent restorations in shadowing (Marslen-Wilson and Welsh, 1978), phoneme monitoring, recognition of word fragments (Grosjean, 1980; Nooteboom, 1981), verbal transformations, intelligibility of lip-reading supported as well as not supported by prosodic cues ('hummed speech') (Risberg, 1979), inferences from historical sound changes (Ohala, 1986).

It would seem that the following must be true for any theory that places the search for invariance exclusively at the level of the signal: Talkers vary their speaking style and thereby contribute to increasing the articulatory and acoustic variability of phonemic and other linguistic units. However, in successful speech acts, such units will always exhibit a core of invariant signal behaviour – absolute or relational – that will remain undestroyed so as to be successfully used by a listener.

Our foregoing review should have made clear why such a stance would run into serious difficulties. We can highlight some of the problems further by means of a very simple example. Consider the following phrase in English:

less'n seven

This utterance can be heard either as 'less than seven', or as 'lesson seven'. In the appropriate contexts (say 'What is the expected number?', and 'What is our topic to-day?') the listener will not be aware of any ambiguity. At which phonetic level do we find the physical correlates of the initial segments of the word 'than'? Needless to say there are no such correlates in this particular case. The conclusion seems inescapable: The quest for invariance – be it absolute or relational – cannot be pursued exclusively at the level of the signal. A more general theory must be sought.

Such a theory might emphasize the following aspects of speech behaviour that have emerged from our review of experimental findings.

The evidence suggests that speech production is shaped primarily by two forces: plasticity and economy. Plasticity is evident when listener-oriented control is called for. Economy is manifest in reductions and other talker-oriented simplifications. These processes interact on a short-term basis so as to generate signals that may be rich or poor in explicit physical information.

We also have evidence to identify two major types of information of perceptual processing: Signal-dependent and signal-independent information. It suggests that on a short-term basis percepts arise from an interaction between the two. Signal-independent can be said to modulate signal-dependent processes in an analogously rich or poor manner.

If we arrange these four components as shown in table I a complementary pattern emerges. When physically poor signals are matched by access to a rich non-signal context, phonetic forms can be successfully derived by the listener. And conversely, when physically rich signals are generated the need for contextual non-signal information is correspondingly reduced.

Table I presents these suggestions in schematized form: Invariance is not an exclusively phonetic phenomenon but can be defined ultimately only at the level of listener comprehension. In emphasizing the complementarity of the components we are in fact raising the question: What would physical speech signals be like, if they were optimal (minimally elaborated), perfectly complementary matches to the listener's running access to signal-independent information? Table I suggests that in the ideal case it is sys-

tematic variability, not invariance, that we should expect to observe at the level of the phonetic signal. The possibility of such complementarity is supported also by some recent results (Hunnicutt, 1985) as well as by an early study by Lieberman (1963).

Table I: Complementarity of perceptual processing mode to signal information contents

SPEECH PRODUCTION MODE	SIGNAL INFORMATION CONTENTS	PERCEPTUAL PROCESSING MODE	COM-PREHENSION
TALKER-ORIENTED	POOR	SIGNAL-IN-DEPENDENT	SUCCESSFUL
LISTENER-ORIENTED	RICH	SIGNAL-DEPENDENT	SUCCESSFUL
	↑ VARIABILITY		↑ INVARIANCE

Needless to say, speakers do not have the mystical power of mind-readers and therefore sometimes fail to make themselves understood and frequently also no doubt produce overly rich forms that overshoot the minimum necessary for being intelligible. Nevertheless, on the basis of the present overview we feel justified in proposing, as a working hypothesis, that intra-speaker phonetic variation is genuine and arises as a consequence of the speaker's adaptation to his judgment of the 'needs of the situation'. In the sense of the biologist's term speech behaviour is an adaptive process.

The proposed way of thinking about the issue does not, of course, rule out finding physical speech sound invariance in restricted domains of observation. For instance, one area where the notion of 'relational invariance' might be productively pursued is inter-speaker phonetic variation exemplified by normalization of individual speaker characteristics. However, our proposal does explain why our quest for a general concept of phonetic invariance has been largely unsuccessful. And, in a pessimistic vein, it predicts in fact that it will continue to be so.

5 On-line processes in the light of typological evidence on phonetic systems

Some time ago Nooteboom did an experiment on word retrieval and was able to show that listeners perform better if presented with the first halves of words rather than with the corresponding second-half fragments (Nooteboom, 1981). To explain his results he suggested that, since word recognition is a real-time left-to-right process, word beginnings are less predictable than word endings. Consequently left-to-right context can be much more easily used than right-to-left context.

He concluded his paper by raising the question whether this asymmetry – that he takes to be a universal feature of the processing of any language – might have left its imprint on how lexical information is organized in the languages of the world. He predicted (p.422) that: *"(1) in the initial position there will be a greater variety of different phonemes and phoneme combinations than in word final position, and (2) word initial phonemes will suffer less than word final phonemes from assimilation and coarticulation rules"*. Evidence indicating that lexicons may at least in part be structured along such lines has been presented by Elert (1970).

We are not in a position to present the typological data needed to test Nooteboom's hypothesis more thoroughly. However, we do have a parallel set of observations that bear on it. (For more detailed discussions see Lindblom and Maddieson (in press), Lindblom, MacNeilage and Studdert-Kennedy (forthcoming)).

This parallel information comes from UPSID, a typological database accessible in a recently published monograph (Maddieson, 1984). It lists consonant and vowel inventories for 317 languages selected genetically and areally so as to form a representative sample of the languages of the world.

The UPSID vowels and consonants were sorted and counted in terms of three categories corresponding to intuitively defined degrees of 'markedness': Basic, elaborated and complex articulations. Basic vowel segments are (short) monophthongs. Elaborations of vowels are introduced when e.g. diphthongization, nasalization, length, retroflexion, pharyngealization, devoicing and other dimensions are invoked. Complex segments are combinations of mechanisms classified as elaborations.

Only three types of segments were observed in UPSID: Systems have basic elements alone, or they consist either of basic and elaborated phonemes, or of basic, elaborated and complex units. Note that this finding implies that there are several combinations of segment types that do not occur. For

instance, there are no inventories with exclusively elaborated or exclusively complex segments. And there are no cases combining complex with basic while omitting elaborated segments. Furthermore, the presence of complex segments presupposes the existence of basic and elaborated phonemes, whereas the converse does not apply. Similarly, if a system has elaborated consonants it will also exhibit basic segments, but the reverse is not found. We conclude from such observations that the data clearly demonstrate a consistent implicational hierarchy among the three types.

Figure 15 shows how vowel systems of the three types are distributed

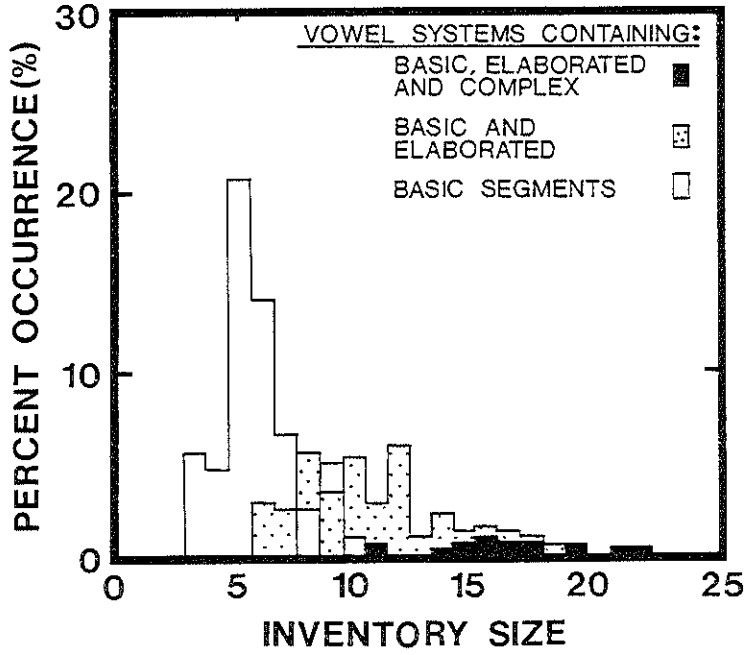

Figure 15: Typological data analysed in terms of a three-way classification of vowel segment types. The data can be interpreted in terms of the 'Size principle'. Small inventories are most frequent. They favour 'basic' vowel articulations. Medium-sized systems invoke 'elaborated' segments in addition. Large inventories are least common. They recruit 'basic' and 'elaborated' vowel types, but also contain 'complex' articulations. Data from 317 languages. Source: Maddieson (1984).

as a function of total size of the vowel inventory. Small systems are the most frequent. They recruit vowels almost exclusively from the basic set.

Medium systems are intermediate in frequency. They invoke elaborated vowels in addition. Large inventories are least common. They bring in complex segments.

Figure 16 illustrates the three-way classification applied to consonants.

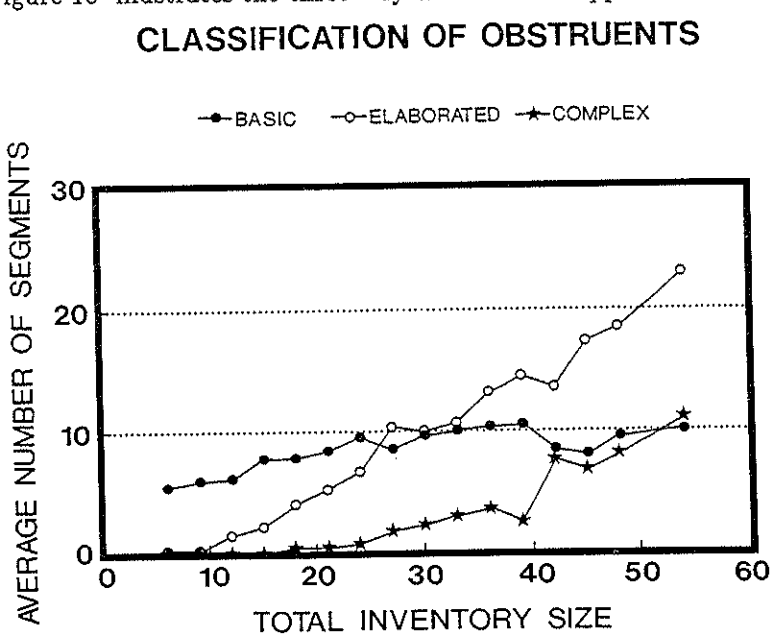

Figure 16: Further evidence of the 'Size principle' as demonstrated by the systematic patterning of consonant types. The number of obstruents, classified as basic, elaborated or complex, is plotted against the total number of consonants per system. The results are similar to those seen for vowels (cf. figure 15). Small inventories favour 'basic' consonant articulations. Medium-sized systems invoke 'elaborated' segments in addition. Large inventories recruit 'basic' and 'elaborated' consonant types, but also contain 'complex' articulations. Averaged data from 317 languages. Source: Maddieson (1984).

The figure shows how basic, elaborated and complex obstruents are invoked as a function of total size of the consonant inventory. Examples of consonant elaborations are features such as breathy, creaky, prenasalized, implosive, pre- and postaspirated, ejective, click. Secondary articulations and extreme places of articulation are also classified as elaborations.

We see from figure 16 that the size is clearly a determinant factor also in the structuring of consonant systems. Small systems recruit consonants

almost exclusively from the basic set. Medium systems invoke elaborated consonants in addition. Large inventories bring in complex segments. Thus it appears that both for vowels and consonants one of the organizing principles underlying the formation of phonetic inventories is inventory size.

It is evident that Nooteboom's expectation is substantiated by these results since his hypothesis in fact implies a 'size principle'. We see the parallel in that a richer, or articulatorily more elaborated, phonetics goes with large vowel and consonant systems in which units are less predictable, on the average, and compete more for distinctiveness. A poorer, or articulatorily less elaborated, phonetics is characteristic of small vowel and consonant systems whose units are, relatively speaking, more predictable. Hence they are exposed to less severe competition for distinctiveness and more open to being structured in terms of articulatory constraints.

We conclude that the phonetic systems analysed exhibit properties that make them functionally adapted to their use. How are the use and the form of language related? What is the short-term phonetic mechanism that underlies the emergence of such adaptations of sound structure? This is a question that John Ohala has dealt with repeatedly in his research and that he will also touch upon in his contribution to this symposium. And it is a question whose answers bear, albeit indirectly, on our present discussion of phonetic invariance.

To show the relevance of this indirect perspective let us first mention a few general conditions that an account of the origin of structure in speech sound inventories must minimally satisfy.

Firstly, the explanation must be mechanistic. Although the patterns revealed by our analysis bear the marks of functional teleological organization we must assume that the mechanism responsible for establishing those patterns is completely blind to its own output. Obviously we do not want to assume that the conscious decisions of language users play a role in the formation of phonological structure. Teleology, or goal-orientation, is an acceptable term describing sound structure but cannot not be invoked with impunity in explaining it.

Secondly, any attempt to account for linguistic facts must conform with the manner in which other historical developments are explained in terms of the principles of general science. In highly abbreviated form a possible argument would run as follows. Since language is a product of biological and cultural evolution, and since the primary mechanism underlying all evolutionary change – whether biological or cultural – is variation and selection (Cavalli-Sforza and Feldman, 1981; Boyd and Richerson, 1986), we are led

to conclude that a model of phonetic variation and selection ought to give us the formalism we need also for explaining how phonology and phonetics are related.

Adopting such a model we consequently find our primary interest naturally focussed on the problem of how to apply the key explanatory concepts of evolutionary theory, variation and selection, to speech and sound structure. A query of primary interest thus becomes: What is the nature of the phonetic variation that feeds the processes of sound system selection? Which conception of on-line speaker-listener interaction is more likely to supply phonologization with the phonetic variation required to create phonetic inventories reported here?

Suppose that we subscribe to a theory of speech which is built on the notion of Signal Invariance – absolute or relative. Please note the implication of adopting such a position. It implies in fact that there is no phonetic variability of linguistic units. There only seems to be because of our presently inadequate conceptual and experimental tools. For, if proved correct, this theory will transform what currently looks like massive variability into observational artefacts. Since it treats variation as an epiphenomenon it will thus offer us little help in accounting for how we are to relate the phonetics of the moment to the patterns fossilized in sound systems.

Our present discussion of the invariance issue has forced us to recognize intra-speaker signal variability as an adaptive and systematic consequence of informational and other functional constraints on natural speaker-listener interactions. When it comes to suggesting explanations for sound pattern regularities this view, let us call it the theory of Adaptive Variability, offers at least two advantages over the theory of Signal Invariance. It treats phonetic variation as genuine. Furthermore, it predicts that this variation ought to be structured in a very specific way, i.e. a way that in fact closely parallels the phonetic organization of sound inventories.

In conclusion, I take the present typological data on phonetic inventories as providing evidence in favour of language structure evolving as an adaptation to the constraints of the on-line processes of speaker-listener interaction and as offering indirect, but strong, additional grounds for adopting the theory of Adaptive Variability, rather than a theory of Signal Invariance, as a heuristic working hypothesis about those processes.

6 Coda: the biology and technology of speech

Our presentation has led us to conclude that the task of resolving the invariance issue is of rather forbidding magnitude although ways of doing so in principle have been suggested. That conclusion may sound pessimistic and must in no way be taken to mean that descriptive acoustic investigations have been shown to be devoid of scientific interest.

Klatt (1977): *"The main objective of future research"* – in the area of speech understanding systems – *"should therefore be the accumulation of more detailed linguistic and acoustico-phonetic facts about English"*. This still holds true for other languages. In a similar vein Fant has repeatedly emphasized that the bottleneck of speech technology is not primarily economic or technical but consists above all in our lack of comprehensive phonetic theory offering us a unified understanding of speech processes (key-note address, International Congress of Phonetic Sciences, Utrecht, 1984).

I completely endorse the message conveyed by those statements and offer my present remarks simply with a view to underlining the magnitude of the task we face in modelling the intricacies of speech biology for applied technological purposes. It took cultural evolution no more than a few centuries to take the step from Leonardo da Vinci's flying machines to present-day aeroplanes. Speech technology has come a long way from the late eighteenth-century 'machine parlante' of Von Kempelen. Also we should bear in mind that, although understanding the biology of behaviour is crucial to engineering applications, after all planes do not flap their wings.

References

Blumstein, S. and Stevens, K.N. (1979) Acoustic invariance in speech production: Evidence from measurement of the spectral characteristics of stop consonants, *Journal of the Acoustical Society of America, 72*, 43-50.

Blumstein, S. and Stevens, K.N. (1981) Phonetic features and acoustic invariance in speech, *Cognition, 10*, 25-32.

Boyd, R. and Richerson, P.J. (1986) *Culture and the evolutionary process.* Chicago: Chicago University Press.

Cavalli-Sforza, L.L. and Feldman, M.W. (1981) *Cultural transmission and evolution.* Princeton, NJ: Princeton University Press.

Cole, R.A. (1973) Listening for mispronunciations: A measure of what we hear during speech, *Perception and Psychophysics, 13*, 153-156.

Delattre, P. (1969) The general phonetic characteristics of languages: An acoustic and articulatory study of vowel reduction in four languages, *Mimeographed*

Report, University of California, Santa Barbara.

Elert, C.C. (1970) *Ljud och ord i Svenskan.* Stockholm: Almqvist och Wiksell.

Engstrand, O. (*to appear*) Articulatory correlates of stress and speaking rate. Accepted for publication in: *Journal of the Acoustical Society of America.*

Fant, G. (1984) Phonetics and speech technology. Keynote address. In: M.P.R. van den Broecke and A. Cohen (eds), *Proceedings of the Tenth International Congress of Phonetic Sciences, Utrecht.* Dordrecht, Holland: Foris Publications, 13-24.

Fischer-Jørgensen, E. (1964) Sound duration and place of articulation, *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, 17*, 175-207.

Flanagan, J.L. (1955) A difference limen for vowel formant frequency, *Journal of the Acoustical Society of America, 27*, 613-617.

Fónagy, I. and Fónagy, J. (1966) Sound pressure level and duration, *Phonetica, 15*, 14-21.

Fowler, C.A. (1986) An event approach to the study of speech perception from a direct-realist perspective, *Journal of Phonetics, 14*, 3-28.

Fowler, C.A., Rubin, P., Remez, R.E. and Turvey, M.T. (1980) Implications for speech production of a general theory of action. In: B. Butterworth (ed.) *Language Production, vol I*, London: Academic Press, 373-420.

Gay, T. (1978) Effect of speaking rate on vowel formant movements, *Journal of the Acoustical Society of America, 63*, 223-230.

Gay, T., Lindblom, B. and Lubker, J. (1981) Production of bite-block vowels: Acoustic equivalence by selective compensation, *Journal of the Acoustical Society of America, 69*, 802-810.

Grosjean, F. (1980) Spoken word recognition and the gating paradigm, *Perception and Psychophysics, 28*, 267-283.

Henke, W.J. (1966) *Dynamic articulatory model of speech production using computer simulation.* Doctoral dissertation, M.I.T.

Hunnicutt, S. (1985) Intelligibility versus redundancy – Conditions of dependency, *Language and Speech, 28*, 47-56.

Keating, P. (1985) Universal phonetics and the organization of grammars. In: V.A. Fromkin (ed.), *Phonetics Linguistics*, Orlando: Academic Press, 115-132

Kelso, J.A.S., Saltzman, E.L. and Tuller, B. (1986) The dynamical perspective on speech production, data and theory, *Journal of Phonetics, 14*, 29-59.

Kewley-Port, D. (1983) Time-varying features as correlates of place of articulation in stop consonants, *Journal of the Acoustical Society of America 73*, 322-355.

Klatt, D.H. (1977) Review of the ARPA speech understanding project, *Journal of the Acoustical Society of America, 62*, 1345-1366.

Kühn, D.P. and Moll, K.L. (1976) A cineradiographic study of VC and CV articulatory velocities, *Journal of Phonetics, 4*, 303-320.

Labov, W. (1972) *Sociolinguistic patterns.* Philadelphia: University of Pennsylvania.

Lacerda, F. (1986) Categories of speech sounds and the dynamics of the auditory system. In: *Speech input/output: Techniques and Applications, IEE Conference Publication, 258*, UK: IEE, 88-93.

Lacerda, F. (1987*a*) Effects of stimulus dynamics on frequency discrimination. In: M.E.H. Schouten (ed.) *The Psychophysics of Speech Perception*, Dordrecht: Martinus Nijhoff, 250-257.

Lacerda, F. (1987*b*) *Effects of peripheral auditory adaptation on the discrimination of speech sounds.* Doctoral dissertation monograph published as *Perilus VI*, Department of Linguistics, Stockholm University.

Lehiste, I. (1970) *Suprasegmentals*, Cambridge: MIT Press.

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

Liberman, A.M., Harris, K.S., Hoffman, H.S. and Griffith B.C. (1957) The discrimination of speech sounds within and across phoneme boundaries, *Journal of Experimental Psychology, 54*, 358-368.

Lieberman, P. (1963) Some effects of semantic and grammatical context on the production and perception of speech, *Language and Speech, 6*, 172-187.

Lindblom, B. (1963) Spectrographic study of vowel reduction, *Journal of the Acoustical Society of America, 35*, 1773-1781 *and* On vowel reduction, Technical report, Department of Speech Communication, RIT, Stockholm.

Lindblom, B. (1967) Vowel duration and a model of lip mandible coordination, *STL-QPSR 4/1967*, Department of Speech Communication, RIT, Stockholm, 1-29.

Lindblom, B., Lubker, J. and Gay, T. (1979) Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation, *Journal of Phonetics, 7*, 147-161.

Lindblom, B., Lubker, J., Lyberg, B., Branderud, P. and Holmgren, K. (1987) The concept of target and speech timing. In: R. Channon and L. Shockey (eds) *In honor of Ilse Lehiste*, Dordrecht, Holland: Foris Publications, 161-182.

Lindblom, B. and Lindgren, R. (1985) Speaker-listener interaction and phonetic variation, *Perilus IV*, Department of Linguistics, University of Stockholm.

Lindblom, B., MacNeilage, P. and Studdert-Kennedy, M. (forthcoming) *Evolution of spoken language.* Orlando: Academic Press.

Lindblom, B. and Maddieson, I. (in press) Phonetic universals in consonant systems. In: L.M. Hyman and C.N. Li (eds): *Language, Speech and Mind*, Croom Helm.

Lindblom, B. and Sundberg, J. (1971) Acoustical consequences of lip, tongue, jaw and larynx movement, *Journal of the Acoustical Society of America, 50*, 1166-1179.

Luce, P.A. (1986) *Neighborhoods of words in the mental lexicon.* Doctoral dissertation, Department of Psychology, Indiana University.

MacNeilage, P. (1970) Motor control of serial ordering of speech, *Psychological Review, 77*, 182-196.

MacNeilage, P. (1980) Speech production, *Language and Speech, 23*, 3-24.

Maddieson, I. (1984) *Patterns of sound*, Cambridge: Cambridge University Press.

Marslen-Wilson, W.D. and Welsh, A. (1978) Processing interactions and lexical access during word recognition in continuous speech, *Cognitive Psychology, 10*, 29-63.

Netsell, R., Kent, R. and Abbs, J. (1978) Adjustments of the tongue and lip to fixed jaw positions during speech: A preliminary report, *Conference on Speech Motor Control*, Madison, Wisconsin.

Nooteboom, S.G. (1981) Lexical retrieval from fragments of spoken words: Beginnings vs endings, *Journal of Phonetics, 9*, 407-424.

Nord, L. (1986) Acoustic studies of vowel reduction in Swedish, *STL-QPSR 4/1986*, Department of Speech Communication, RIT, Stockholm, 19-36.

Ohala, J.J. (1986) Phonological evidence for top-down processing in speech perception. In: J. Perkell and D. Klatt (eds) *Invariance and variability in speech processes.* Hillsdale, NJ: Lawrence Erlbaum Associates, 386-401.

Ohala, J.J. and Feder, D. (1986) Speech sound identification influenced by adjacent 'restored' phonemes. Paper ZZ13 presented at the 112th meeting of the ASA, *Journal of the Acoustical Society of America, 80*, S110.

Öhman, S. (1966) Coarticulation in VCV utterances: Spectrographic measurements, *Journal of the Acoustical Society of America, 39*, 151-168.

Öhman, S. (1967) Numerical model of coarticulation, *Journal of the Acoustical Society of America, 41*, 310-320.

Perkell, J. and Klatt, D. (1986) *Invariance and variability in speech proccesses.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Pollack, I. and Pickett, J.M. (1964) Intelligibility of excerpts from fluent speech: Auditory vs structural context, *Journal of Verbal Learning and Verbal Behaviour, 3*, 79-84.

Risberg, A. (1979) *Bestämning av hörkapacitet och talperceptionsförmåga vid svåra hörelskador.* Doctoral dissertation, Royal Institute of Technology, Stockholm.

Schulman, R. (forthcoming) Articulatory dynamics of loud and normal speech. Submitted to *Journal of the Acoustical Society of America.*

Stevens, K.N. and House A.S. (1963) Perturbation of vowel articulations by consonantal context: An acoustical study, *Journal of Speech and Hearing Research, 6*, 111-128.

Stevens, K.N. and Blumstein, S. (1978) Invariant cues for place of articulation in stop consonants, *Journal of the Acoustical Society of America, 64*, 1358-1368.

Stevens, K.N. and Blumstein, S. (1981) The search for invariant correlates of phonetic features. In: P. Eimas and J. Miller (eds) *Perspectives on the study of speech*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Sundberg, J. (1975) Formant technique in a professional singer, *Acustica, 32*, 89-96.

Traunmüller, H. (1981) Perceptual dimension of openness in vowels, *Journal of the Acoustical Society of America, 69*, 1465-1475.

Traunmüller, H. (1985) The role of the fundamental and higher formants in the perception of speaker size, vocal effort and vowel openness, *Perilus IV*, Stockholm University, 92-102 .

Warren, R. (1970) Perceptual restoration of missing speech sounds, *Science, 167*, 392-393.

Westbury, J. and Keating, P. (1980) Central representation of vowel duration, *Journal of the Acoustical Society of America, 67*, Suppl. 1, S37(A).

Discussion of Björn Lindblom's 'Phonetic Invariance and the Adaptive Nature of Speech'

John J. Ohala*

1 Introduction

I compliment Björn Lindblom on a masterful well-documented presentation that consitutes, I think, a conceptual breakthrough in the troublesome problem of invariance. He has boldly presented a position which challenges prevailing wisdom and the foundations of some well-known research paradigms. The existence of clearly-defined competing hypotheses will promote well-focussed research which will help to resolve this conflict.

His account of the origin of speech variability is explicitly a biological one and uses terms and concepts common in biology such as 'adaptation' and 'evolution'. This is not just a metaphor; he quite correctly treats speech as a biological activity and believes that its shape and behaviour must be governed by and therefore explainable by biological principles, i.e. by reference to physical and physiological constraints impinging on speech (see also Zipf, 1935).

I am convinced by much of his argument and will offer here a few brief remarks which supplement some of the points he made, in particular the connection between the biologist's and linguist's treatment of variation. In a few instances, I will suggest some qualifications to his scheme.

2 Parallels between biology and linguistics

Preliminaries

Oversimplifying a bit, the biologist must explain two types of variability: the 'plastic' behaviour and forms of organisms in the face of differing environmental pressures, e.g. temperature regulation by means of perspiration,

* Phonology Laboratory, Department of Linguistics, University of California, Berkeley, California 94720, USA.

WORKING MODELS OF HUMAN PERCEPTION
ISBN 0-12-238050-9

adjusting the thickness of fur or feathers – this can be called short-term adaptation – and the variable genetically-determined shapes and behaviours of organisms over time due to natural selection – this is long-term adaptation. The first may be characterized as purposeful or teleological in that it is controlled by the error between a pre-set goal and the detected current state of an organism. Variable behaviour of this type can be continuous as it is, for example, in the case of thermoregulation. Lindblom's adaptive variability is of this type, where, as the evidence he has reviewed shows, the variable behaviour is determined by the speaker's estimation of the decoding capabilities of the listener (see Nooteboom(1983) for further arguments and evidence towards this point). The second type of variation in biology, caused by natural selection, is not purposeful or teleological; it operates blindly. More adaptive organisms simply propagate better than less-well-adapted ones. Does this have an analogue in speech? Before committing ourselves on this point, let us take a closer look at the mechanisms Darwin proposes which underlie natural selection.

Darwinian evolution

Briefly, Darwin's theory explains the origin of species on the basis of three principles:

1. There exists 'natural' genetic variation in offspring vis-à-vis their parents. This is due to the random shuffling of the genes from both parents in the case of sexual reproduction and due to imperfect copying or other distortion (mutation) of the genetic code in sexual or asexual reproduction.
2. The demand for resources needed to sustain life (territory, food) exceeds the supply.
3. Those individuals whose variable genetic endowment ('genotype') makes them more competitive in garnering the available resources will reproduce in greater numbers than other less competitive.

From these axioms is derived the main theory of evolution that, in time, these adaptive variations will breed true and will no longer be random; different species will result. Genetically-maintained variations differ from the plastic adaptations not only in being non-teleological in origin but also by being discrete. An okapi is born with a definite limit on the extent to which it can stretch its neck; if it wants to stretch its neck further, it has to become its cousin, the giraffe!

Sound change

It would seem that the closest analogue in speech to the variation that gives rise to different species is *sound change*. Sound change, of course, is the change in pronunciation norms from one generation to the next, e.g. the differences between Chaucer's 14th century pronunciation [wɪf], 'wife', and the current English pronunciation of [waɪf]. When a sound change affects one regional or social speech community but not another, it can give rise to dialect differences, e.g. British and American English [streɪt] 'straight' vs Australian [strʌɪt] or, of course, to different languages. When it affects a given morpheme in one phonological context but not in another it can lead to alternation such as [lɔŋ], 'long', and [lɛŋkθ], 'length', or to complete merger and homophony as when 'less than' merges with 'lesson' in the example Lindblom cited (similar examples of homophony abound and constitute the stuff out of which jokes and puns are made). Sound change, as he mentioned, can also affect a language's segment inventory: if the sound change causes a complete merger of two previously distinct sounds the inventory is reduced, as has happened to Western American English in the case of the vowels [ɔ] and [ɑ], leading to homophony of words such as 'caught' and 'cot'. Augmentation of segment inventories also occurs when what were once predictable phonetic variants of a given sound become distinctive or non-predictable, as happened in the history of English some nine centuries ago: the previously predictably voiced variants of fricatives – occurring only in intervocalic position – became distinctive, thus leading to morpheme alternations such as 'wife' – 'wives', 'waft' – 'wave', etc.

Differences between sound change and biological evolution

Although sound change is similar in many respects to the evolution of species, I think there are some subtle but important differences between the two phenomena. In what follows, I present a brief summary of my own work on the mechanisms of sound changes. These matters are controversial, but I have attempted to provide empirical support for my claims by reference to a variety of laboratory studies (see Ohala, 1974, 1978, 1981, 1983*abc*, 1985, in press; Ohala and Lorentz, 1977; Ohala and Riordan, 1979). Current accounts of the causes of sound change, like the theory of evolution (of species), assume the existence of natural variation in pronunciation[1].

[1]It must be recognized that pronunciation may change due to a multiplicity of factors, some of them non-phonetic, e.g., paradigm regularization, spelling pronunciations, etc.

Some of this variation becomes 'fixed' or lexicalized into sound changes. What is the source of this variation? It may be possible, as Lindblom claims, that some of this variation occurs due to the kind of speaker adaptation he has described and that some effects of sound change thus represent a 'fossil record' of adaptive variation, but this has not yet been demonstrated empirically. I rather think that the majority, if not all, of such variation – like the variations in genotypes – is mechanical and non-purposeful.

There is good evidence for this latter view. First, some variation in pronunciation almost certainly is not under the full control of the speaker but rather crops up because of physical constraints of the speaking mechanism. For example, it is aerodynamic constraints that give rise to noisy releases of stops before high close glides and vowels. This can be misconstrued by listeners as intentional and lead to sound changes such as [ækt ʃuəl] 'actual' from [ækt+juəl]. The fortuitous character of some variations seems to be demonstrated by the fact that listeners usually factor them out: Mann and Repp (1980) showed that listeners show a crossover between [s] and [ʃ] at a lower frequency when the following vowel is [u] in comparison to following [ɑ], presumably because they are aware that the assimilated labial rounding during the [s] would fortuitously lower its centre frequency (for other examples, see Ohala, 1981; Beddor, Krakow and Goldstein, 1986; Ohala and Feder, 1987). (Of course, it must not be assumed that listeners always succeed in factoring out such distortions in speech. Indeed, the small fraction of cases where they fail can also be a source of sound change; see below). Furthermore, a significant fraction of the variation can be shown to occur not in the mouth of the speaker but in the ears of the listener (Sweet, 1900, p.21-22; Jonasson, 1971; Ohala, 1981, 1983*ab*) through what must be innocent misapprehensions, that is, neither purposeful nor adaptive in any sense. For example, labialized velars [kʷ gʷ xʷ] are confused with simple labial consonants [p b f], thus giving rise to Modern English [læf] 'laugh' with [f] where the conservative spelling reveals there was once a velar fricative [χ] (labialized by virtue of the preceding labial glide). Also, a completely new series of consonants, e.g. the palatalized series in Slavic languages, can presumably arise because of what may be called 'parsing' errors by the listener: a mistake in assigning the 'sharpness' (high F_2 transition) to an adjacent consonant rather than to a vocalic segment.

However, if we focus on sound changes of the type that are found in many different languages, though they be typologically, geographically, chronologically, and genetically separated, we can be fairly sure we are dealing with those caused by universal physical factors.

That it is listeners' misapprehensions which underlie a major fraction of sound changes is evidenced by the fact that listeners' confusions in lab-based listening tests parallel sound changes to a degree that cannot be due to chance: in the nature of the change, in the environments that promote it, in the asymmetrical directionality of the change (if any). For example, in both sound change and listening tests we find [p] changed to (confused with) [t] primarily in the environment of following [i] (or if the labial stop is palatalized); in addition, this change/confusion is asymmetrical since [t] rarely changes to [p] (Ohala, 1978, 1983*a*).

To come back to the comparison with natural selection, then, I have suggested that the evolution of pronunciation, sound change, is similar in that it starts with a kind of natural variation, but is dissimilar in that it makes no assumption that there is substantial ecological competition between pronunciation norms or that most variants are any more adaptive than others. I should clarify this last statement since it may seem to contradict an earlier observation that some sounds or sound combinations are more subject to confusion than others. The point is that physical (including physiological) principles constrain what variants come into being by influencing the fortuitous distortions in the speech of speakers and the hearing and articulatory capabilities of misapprehending listeners, but after those variants exist there is little evidence for subsequent optimization through competition of languages' sound systems. (Prague School phonologists would not agree with this, cf. Jakobson, 1972 [1931]). I think most variations occur due to errors in the transmission of pronunciation norms – due to listeners' mistakes – and thus resemble scribes' errors in copying manuscripts. Like scribal errors, there is no adaptive value to such variations. Further support for this view comes from the observation that over the time span that linguists have been able to investigate the history of languages, c. six millennia, – in which time many languages' phonologies, including their segment inventories, have undergone remarkable changes – there has been no detectable improvement in the communicative capacity of speech.

Now I must qualify this claim by admitting that, although the variations that become fossilized in sound change do not *per se* make speech adaptive, it may nevertheless be true that if one variant is adopted by or associated with a prestigious speaker or group, that might make it propagate more widely than others. This is a sociolinguistic fact: pronunciation norms may be adaptive (benefit the speaker) if they impart some desired social status. This dimension, however, really lies outside the strictly phonetic domain I have been discussing and so does not contradict my claim.

On Lindblom's analysis of universal tendencies in segment inventories

For this reason, I am not fully convinced that Lindblom's analysis of the Maddieson data, where large segment inventories show more 'marked' segments than small ones, necessarily represents a fossilization of adaptive variation, i.e. the purposeful variation speakers implement as a function of listeners' needs. It is as likely, I think, that such augmentation occurs mechanically and inadvertently due, e.g., to listeners' errors in parsing the speech signal. We should also keep in mind that the addition of 'marked' consonants does not necessarily increase the overall distinctiveness of consonants in proportion to their added number. Put more explicitly: the increment of distinctiveness between one unmarked segment and another, e.g., [p] and [b], is greater than the distinctiveness between an unmarked segment and the marked segment that developed from it, e.g., [p] and [p']. There is evidence that the palatalized consonants in Russian (which count as 'marked') are much more subject to confusions than are the plain (unmarked) consonants under various listening conditions (A. Stern, personal communication).

Speech style not necessarily correlated with distinctiveness

These points also have some bearing on the *discrete* style-dependent phonological recodings of pronunciation mentioned by Lindblom, e.g., 'native' as [neitʰɪv] in more formal, careful speech or [neiɾɪv] in more casual style. The existence of these alternants is due to sound change; the one form is historically derived from the other but is no longer so derived by the speaker; they are, as Lindblom implied, just style-dependent variant forms each stored separately in the speaker's lexicon. Not all such alternants correlate with greater vs lesser distinctness as a function of the listener's greater vs lesser need for it, respectively. The style-dependent variants of the English present participle suffix [-ɪŋ] (formal) and [-ɪn] (informal) do not, as far as I know, differ in terms of distinctness.

3 The lesson of adaptive variability

The notion of adaptive variability is an important one for many practical tasks facing us currently, e.g., trying to construct robust devices for the automatic speech recognition (ASR). I have tried to strengthen Lindblom's

claim by removing questionable evidence offered in support of it. If sound change has no direct connection to adaptive variability does it nevertheless have some relevance to tasks such as ASR? I think it does, in a number of ways. First, given the vast treasure of data on sound change which linguists have accumulated in a little less than two centuries, it provides us with valuable information on the favoured paths of speech sound variation and confusion. Where human ears have stumbled (apologies for the incongruous metaphor) there will the ASR device stumble, too (see Ohala 1975, 1983*b*, 1985, 1986*a*). Second, a close examination of the record of sound changes, in particular those known as 'dissimilation', shows that the listener is not passive in the process of speech communication; the speaker is not the only one to 'adapt' to the constraints of the speaking situation. Dissimilation is the process whereby two sounds having similar features undergo change such that the shared feature is removed from one of them. An example is Latin [kʷTŋkʷə], 'five', with two labialized velars which became [kTŋkʷə] with the labialization removed from the first velar. (Subsequently this first velar changed to an affricate or fricative, giving Italian [tʃiŋkʷə], Spanish [siŋko], etc.). I have suggested (Ohala, 1981, 1985, 1986*a*) that this came about due to the listener erroneously attributing the labialization on the first velar to the fortuitous spillover of the labialization from the second velar and therefore eliminating it from his own most careful pronunciations. This is an auditory analogue of the visual error we call 'camouflage'. Crucial to this analysis is the idea that the listener tries to 'make sense' of the speech signal by 'parsing' it (see also: Beddor *et al.*, 1986; Ohala, 1986*b*). Undoubtedly the listener does this successfully in most cases, but occasionally makes mistakes that can show up as the sound change dissimilation. Third, a study of linguists' pronouncements on sound change shows that we should be quite cautious in trying to identify those things in speech that we expect to be the 'same'. Linguists are good at identifying the common historical origins of different words, morphemes, sounds – for example, they can show that the roots of '*eques*trian' and '*hippo*potamus' are the 'same' (both from the Indo-European word for horse **ekwōs*, but it is not clear that this notion of 'same' is the one that will do much good in an ASR task. The sameness of the roots in 'equestrian/hippopotamus' is an extreme case, perhaps, and not really one that would mislead anyone engaged in ASR. But there are other, more troublesome cases, e.g., the alleged 'sameness', according to linguists, of the variants of the phoneme /t/ in English, namely [tʰ], [t], [ɾ] and [ʔ] in [tʰɑp], 'top', [stɑp], 'stop', [bɛɾi], 'Betty', and [mɑ̃ʔn̥], 'mountain', respectively. Are these the 'same' sounds or simply different sounds that

had a common historical origin? We would do well to keep in mind that the linguists know historical 'sameness' best and have not yet refined their techniques to identify psychological or neuromotor sameness (however, see Jaeger, 1980, 1986; Ohala, 1983*a*; Derwing and Nearey, 1986).

4 Conclusion

In summary, Lindblom has presented a convincing case that invariance is an elusive concept in speech and is not to be found in the articulatory, acoustic, or even in the auditory domain; rather, it is something that must be constructed by the listener using whatever clues are available, even non-phonetic ones.

References

Beddor, P.S., Krakow, R.A., and Goldstein, L.M. (1986) Perceptual constraints and phonological change: A study of nasal vowel height. *Phonology yearbook 3*, 197-217.

Derwing, B.L. and Nearey, T. (1986) Experimental phonology at the University of Alberta. In: J.J. Ohala and J.J. Jaeger (eds) *Experimental phonology*, Orlando: Academic Press, 187-209.

Jaeger, J.J. (1980) Testing the psychological reality of phonemes. *Language and Speech, 23*, 233-253.

Jaeger, J.J. (1986) Concept formation as a tool for linguistic research. In: J.J. Ohala and J.J. Jaeger (eds) *Experimental phonology*, Orlando: Academic Press, 211-237.

Jakobson, R. (1972) Principles of historical phonology (Translation by A.R. Keiler of original 1931 German). In: A.R. Keiler (ed.) *A reader in historical and comparative linguistics*, New York: Holt, Rinehart and Winston, 121-138.

Jonasson, J. (1971) Perceptual similarity and articulatory reinterpretation as a source of phonological innovation. *Speech Transmission Laboratory Quarterly Progress and Status Report, 1/171*, Stockholm, 30-41.

Mann, V.A. and Repp, B.H. (1980) Influence of vocalic context on perception of the [š] vs [s] distinction. *Perception and Psychophysics, 28*, 213-228.

Nooteboom, S.G. (1983) Is speech production controlled by speech perception? In: M. van den Broecke, V. van Heuven and W. Zonneveld (eds), *Sound structures. Studies for Antonie Cohen*, Dordrecht: Foris Publications, 183-194.

Ohala, J.J. (1974) Experimental historical phonology. In: J.M. Anderson and C. Jones (eds) *Historical linguistics II. Theory and description in phonology*, Amsterdam: North Holland, 353-389.

Ohala, J.J. (1975) How a study of sound change can aid in automatic speech recognition. In: G. Fant (ed.) *Speech Communication, volume 3: Speech perception and automatic recognition*, Stockholm: Almqvist and Wiksell, 299-302.

Ohala, J.J. (1978) Southern Bantu vs the world: the case of palatalization of labials. *Berkeley Linguistic Society, Proceedings of Annual Meeting 4*, 370-386.

Ohala, J.J. (1981) The listener as a source of sound change. In: C.S. Masek, R.A. Hendrick, and M.F. Miller (eds) *Papers from the Parasession on Language and Behavior*, Chicago: Chicago Linguistic Society, 178-203.

Ohala, J.J. (1983*a*) The phonological end justifies any means. In: S. Hattori and K. Inoue (eds) *Proceedings of the XIIIth International Congres of Linguists, Tokyo, 29 August - 4 September 1982*, Tokyo: Sanseido Shoten, 232-243.

Ohala, J.J. (1983*b*) Modern applied linguistics. In: *Proceedings of the Arab School on Science and Technology, First Fall Session: Applied Arabic Linguistics and Signal and Information Processing*, Damascus, Syria, 51-62.

Ohala, J.J. (1983*c*) The direction of sound change. In: A. Cohen and M.P.R. van den Broecke (eds) *Abstracts of the 10th International Congres of Phonetic Sciences*, Dordrecht: Foris Publications, 253-258.

Ohala, J.J. (1985) Linguistics and automatic speech processing. In: R. De Mori and C.-Y. Suen (eds) *New systems and architectures for automatic speech recognition and synthesis*, (NATO ASI Series, Series F: Computer and System Sciences, Vol. 16), Berlin: Springer-Verlag, 447-475.

Ohala, J.J. (1986*a*) Phonological evidence for top-down processing in speech perception. In: J.S. Perkell and D.H. Klatt (eds) *Invariance and variability in speech processes*, Hillsdale, NJ: Lawrence Erlbaum Associates, 386-397.

Ohala, J.J. (1986*b*) Against the direct realist view of speech perception. *Journal of Phonetics, 14*, 75-82.

Ohala, J.J. (in press) The phonetics and phonology of aspects of assimilation. In: a volume edited by M. Beckman and J. Kingston.

Ohala, J.J. and Feder, D. (1987) Listeners' identification of speech sounds is influenced by adjacent 'restored' phonemes. In: *Proc. 11th International Congress of Phonetic Sciences, August 1-7, 1987, Tallinn, Estonia, USSR, Vol. 4*, 120-123.

Ohala, J.J. and Lorentz, J. (1977) The story of [w]: an exercise in the phonetic explanation for sound patterns. *Berkeley Linguistic Society, Proceedings of Annual Meeting 3*, 577-599.

Ohala, J.J. and Riordan, C.J. (1979) Passive vocal tract enlargement during voiced stops. In: J.J. Wolf and D.H. Klatt (eds) *Speech communication papers*, New York: Acoustical Society of America, 89-92.

Sweet, H. (1900) *The history of language*. London: J.M. Dent.

Zipf, G.K. (1935) *The psycho-biology of language*. Boston: Houghton Mifflin Co.