# Coarticulation and the so-called 'Invariance Problem'

Sven Öhman
Department of Linguistics
Uppsala University

A widespread conception of the phenomenon of coarticulation may be captured by the following definition:

**Coarticulation** is the simultaneous realisation in connected speech of distinct, consecutive phonological units (phonemes, segments).

A typical example of this phenomenon is encountered in common pronunciations of the English word *two*. The aspirative release of the initial voiceless plosive [t] is audibly colored by the lip rounding of the following vowel [u]. You might therefore say that the initial [t] is *coarticulated* with the final [u]. The [t] and the [u] are "simultaneously realised" as stated in the definition above in that there is no sharp acoustic boundary between the [t]-quality and the [u]-quality.

Phoneticians and speech technologists tend to feel that the coarticulation schema just mentioned points to a fact of considerable generality, *viz.* that phonological units (phonemes, segments) are in general 'not represented by invariant speech outputs'. For if we put our initial [t]-consonant into a different phonological environment, say into that of the word *tea*, for example, the influence of the vowel with which the [t] is coarticulated this time is rather different from that suffered in the word *two*.

In general we may describe this situation with the statement that definitions of phonological units (phonemes, segments) in terms of necessary and sufficient conditions on corresponding acoustic events of the speech output are not feasible in any exhaustive way.

This summarises what is commonly called *the 'invariance problem'*.

Needless to say the invariance problem is a serious obstacle to making the 'speech type-writer' come true or, in fact, to any attempt at converting speech into text by machine, at least if one tries to bring this conversion about by detecting consecutive phonological units in the acoustic speech output.

I will not try to recapitulate here the many engineering methods to this end that have been found to fail throughout the half-century in which speech research has so far existed.

What I want to propose is, on the contrary that the main problem we are facing here is *not* of a technological nature, but that it is of a *purely conceptual* kind. I. e., the real problem is mainly due to a misunderstanding of the very essence of our problem. This essence is *not* the difficulty, or even the impossibility, of setting down the necessary and sufficient conditions on the acoustic events of the speech output that supposedly correspond to some phonological units whether phonemes or segments. It is rather that the conventional wisdom of contemporary linguistics including phonology makes us blind to the reasons why it is absurd to expect that there should be any such 'necessary and/or sufficient conditions' in the first place!

A whole army of terminological conventions stand in the way of our seeing things clearly here. Thus tape recorded speech is called 'speech output', suggesting that there is some kind of 'input' that somehow triggers this alleged 'output' whenever an ordinary human being is saying something!

Obviously, this strange language comes from taking speech synthesis too seriously as a model for human speech production, instead of seeing what it really is - a gadget to generate (at best) speech-like sounds.

This curiously mechanistic-mentalistic picture of natural speech as the outputs reflecting phonological 'mind-inputs' gets a sort of backing from the metaphysics of Chomsky-Halle type psychological realism according to which the linguist's phonological representation of a stretch of speech miraculously has an existence as an independent 'mental reality' in the speaker/hearer's mind even before he or she ever had any experience of reading or writing or of the way linguists think or theorise about language.

We find the metaphysics of phonological 'input' - acoustic speech 'output lurking already in the terminology involving the words *realise* and *represent.* Thus in my above quoted definition of coarticulation this phenomenon is said to involve the simultaneous *realisation* of distinct, consecutive phonological units (phonemes, segments).

This language conjures up the (false) idea that the alleged phonological units (phonemes and/or segments) have a real existence as *potentialities* of a kind not unlike actual objects (presumably residing somewhere in 'the mind of the speaker/hearer) 'mental objects', that is, which are made to materialise in the form of audible sound in actual human speech.

The phonological units (phonemes, segments) do not exist, however, outside the linguists' or other people's phonological representations of actual or possible expressions of language. They are means of systematically and methodically depicting speakable utterances as it were, methods of representation laboriously worked out by specialists along with techniques of use that must be laboriously trained in and practically mastered.

We may find some methods of phonological representation more convenient or natural than others, but we deceive ourselves if we try to believe that such advantages be due to the representation technique somehow having a life of its own *in* the utterances that we represent by its means!

Along with the techniques of representation the student also acquires a way of seeing the facts of language, a way of seeing which is so-to-speak inherent in the mastery of the techniques. It is by hypostasizing this way of seeing that one comes to consider the units of phonological representation (phonemes, segments) as a kind of autonomous (e. g. mental) entities. If we are to say that they are in any way *realised* we should presumably say that they are realised in the linguist's concretely representing some actual or possible utterance of speech, but then only in the sense that he actually writes down the phonological units in question.

The other side of the so-called 'invariance problem' has to do with the by now notorious 'speech output'. And here it is of the utmost importance that we make clear to ourselves *how* we know anything about this phenomenon.

In contemporary phonetics the most widely practised method of *recording* speech is of course that of taping it. Digitalised computer stored speech is of course also available to us although I believe that this method of recording is usually obtained by passing through a previous tape recording.

These techniques would of course be worthless to the student of human speech unless it were possible *to listen* to the recorded speech by playing it back. We ought therefore to say, perhaps, that tape- and/or computer recording-*plus* -listening is the most important way in which a contemporary phonetician comes to acquire direct knowledge of speech.

By using the available play-back facilities of tape recorders and computers the phonetician can observe individual acoustic events that take place in a recording. But we can of course also obtain a *representation* of these events in the form of a so-called sound spectrogram (e. g. of the SonaGraph type).

Needless to say, our understanding of this kind of representation is intimately dependent on our understanding of the conceptual context within which the spectrogram has its natural place - i. e. our understanding of physical phonetic-acoustics and of its electronic- acoustic technology.

When we study the facts of coarticulation and worry about the so-called 'invariance problem' we thus have in front of us, lying side-by-side, two essentially different kinds of representation of one and the same spoken utterance - one is a physical-acoustic representation comprehensible only in light of the natural science concepts and techniques that come along with it - and, the other, a linguistic-phonological representation with *its* concepts and techniques - and there is naturally enough no necessary logical link between the two!

There is simply no reason to expect that there should be any direct (or indirect) connexion between them - *of and by itself*!

If we want there to be such a connexion, we have to invent it.

Building a machine that reacts in various ways to sounds of speech and then learning to talk in such a way that the machine reacts in the desired ways is the recipe for such an invention.

The speech sounds required by the machine for proper action may of course vary a lot and need not resemble ordinary human speech very much although it might do so - after a sufficiently long period of experimentation.

Also, we should not take it for granted that people will be unwilling to take advantage of even quite unspeech-like 'speech type-writers for, their value to the customer will most probably depend on what you can do with them - what kind of tasks they will let you accomplish. Being required to grunt or to groan might not be totally objectionable to everyone*under all circumstances*!

One may imagine the coming about of this kind of speech machine on a large scale in all or most turns of life. Our whole language culture would then certainly be considerably altered.

Will we be willing to go along with all this? That depends, I believe, on what, by way of increased life possibilities, we will gain from it.